

September 2004 Draft  
For Discussion and Feedback

**Building Credibility into Performance Assessment and  
Accountability Systems for Teacher Preparation  
Programs:**

**A “How To” Manual for Teacher Educators Who Want  
to Collect, Use and Report Valid and Reliable  
Performance Data on Teacher Candidates with A Link  
to P-12 Student Learning**

*Developed by:*

**Peter Denner, Idaho State University  
Roger Pankratz, Western Kentucky University  
Tony Norman, Western Kentucky University  
Jack Newsome, Idaho State University**

**A Work in Progress Sponsored by the Renaissance Partnership  
For Improving Teacher Quality and Supported by a  
Title II Teacher Quality Enhancement Grant**

# Introduction

The processes described in this manual have been developed through four years of experiences at eleven universities in the Renaissance Partnership that prepare about 6,000 teachers each year. The eleven universities located across the country, from California to Michigan to Pennsylvania and to Tennessee, have been part of a Title II Teacher Quality Enhancement Grant Program since 1999. Key project objectives include:

1. Developing fully operational performance assessment and accountability systems that meet NCATE 2000 Standard 2.
2. Introducing and using teacher work samples as instructional and assessment tools that link teacher performance to P-12 student learning.
3. Mentoring teacher candidates to design and teach high-performing instructional units based on state and local content standards, assess student learning, report learning results and evaluate their own teaching and learning processes.

This manual is the result of the Renaissance Partnership's collective experiences with Objective 1 above, a most difficult challenge faced by all teacher preparation institutions: building credibility into performance assessments of teacher candidates and developing data management systems that can produce user-friendly reports for candidates, faculty, program administrators, accrediting agencies, employers and policy makers.

At this point in our development, the Renaissance Partnership members have identified and used twelve processes to establish "validity" and "generalizability" (scoring consistency) for measures of teacher candidate performance and designing functional data management systems. These include:

1. Examining existing measures of teacher candidate performance
2. Designing performance assessments that have a high potential to measure identified teaching standards fairly and consistently
3. Producing teacher candidate performances for validity and reliability studies
4. Planning credibility evidence studies for performance assessments
5. Collecting candidate performances for credibility evidence studies
6. Training raters to score performances
7. Benchmarking performances
8. Scoring performances
9. Determining score generalizability
10. Gathering validity evidence
11. Developing your data collection timeline & initial reports
12. Examining the consequential validity of your assessment system

These twelve processes are those we have worked through and found useful in our quest to design fully operational accountability systems at each of our eleven universities. The developers of this manual would like to share with other institutions and teacher educators our experience and work. The examples provided in the manual are based on teacher work samples and accountability systems being developed at Western Kentucky University and Idaho State University. In the future, we hope to show how the processes have universal applications for all teacher performance assessments and teacher preparation accountability systems.

However, at this point in our work we have focused on examples from our immediate experiences.

We hope you will examine and use the processes in this manual and provide us feedback. Additional information and resources on our work and the Renaissance Partnership Project can be accessed by going to our web site <http://fp.uni.edu/itq>. You can best communicate your comments and suggestions about one or more parts of this manual to any of the authors via e-mail.

*Peter Denner* [dennpete@isu.edu](mailto:dennpete@isu.edu)  
*Roger Pankratz* [roger.pankratz@wku.edu](mailto:roger.pankratz@wku.edu)  
*Tony Norman* [antony.norman@wku.edu](mailto:antony.norman@wku.edu)  
*Jack Newsome* [newsjack@isu.edu](mailto:newsjack@isu.edu)

Draft

## TABLE OF CONTENTS

Component 1: Examining existing measures of teacher candidate performance .....	5
Component 2: Designing performance assessments that have a high potential to measure identified teaching standards fairly and consistently .....	11
Component 3: Producing teacher candidate performances for validity and reliability studies ....	21
Component 4: Planning credibility evidence studies for performance assessments .....	24
Component 5: Collecting candidate performances for credibility evidence studies .....	27
Component 6: Training raters to score performances .....	29
Component 7: Benchmarking performances .....	32
Component 8: Scoring performances .....	35
Component 9: Determining score generalizability .....	37
Component 10: Gathering validity evidence .....	42
Component 11: Developing your data collection timeline & initial reports .....	52
Component 12: Examining the consequential validity of your assessment system .....	62

## **Component 1: Examining Existing Measures of Teacher Candidate Performance**

**Key Question:** *How do we currently measure the knowledge, skills, and dispositions of our teacher candidates, and what additions/revision do we need to make?*

**Task for this Component:**

Examine the degree to which existing measures of candidate assessment for program completion address all required national, state, and local teaching standards and identify specific areas where performance assessments need to be developed or improved.

**Rationale for this Component:**

The first step in assuring that the knowledge, skills, and dispositions of teacher candidates are assessed in a valid, fair, and consistent manner is to examine the ways in which performances are currently assessed. Once candidate performance standards are established, evidence should be gathered to confirm that (a) assessment measures match well with the methods used to develop the skills, in both content and cognitive complexity; (b) every standard is assessed adequately; (c) all assessment measures are administered and scored consistently; and, (d) candidates have multiple opportunities to demonstrate teaching standards. The evidence you gather will enable you to plug any gaps and address any weaknesses in your assessment system prior to expending valuable time and resources collecting credibility data on candidate performance assessment measures.

**Process for this Component:**

**Step 1. Develop a Matrix to Assess the Adequacy of Existing Assessment Measures**

Develop a matrix that will enable you to review the adequacy of existing performance measures and their alignment with performance standards. List the teacher standards your program must address down the left column and your existing performance measures across the top. Table 1 presents an example for Western Kentucky University (WKU) that identifies nine Kentucky new teacher standards and four existing instruments used to assess teacher performance candidates for exit from the preparation program. There are four existing teacher candidate performance instruments used at WKU: (1) an observation record, (2) a teacher work sample, (3) PRAXIS II and (4) a professional development plan (PDP) submitted in a portfolio. A space is provided in each column to judge the “extent of coverage” and “quality of measurement” of each instrument relative to each standard.

A fifth column is provided to complete a summary assessment of what needs to be considered for development or improvement of existing measures.

**Table 1. Teacher Standards versus Assessment Measure Matrix (WKU Example)**

Kentucky Teacher Standards	Existing WKU Assessment Measures								Summary of Development and/or Improvement Needs
	Classroom Observation Record		Teacher Work Sample		Praxis II		PDP Plan Portfolio		
	Standard Coverage	Instrument Quality	Standard Coverage	Instrument Quality	Standard Coverage	Instrument Quality	Standard Coverage	Instrument Quality	
<b>Standard I</b> Knows Teaching Content									
<b>Standard II</b> Designs Instruction									
<b>Standard III</b> Creates Learning Environment									
<b>Standard IV</b> Implements Instruction									
<b>Standard V</b> Assesses Learning									
<b>Standard VI</b> Reflects and Evaluates Teaching and Learning									
<b>Standard VII</b> Collaborates with Parents and Professionals									
<b>Standard VIII</b> Engages in Professional Development									
<b>Standard IX</b> Uses Technology for Instruction									

## **Step 2. Review Existing Performances for Adequacy of Coverage and Measurement Quality**

Use the symbols H, M, L or O, respectively, for High, Medium, Low or no coverage or no existing instrument. Make a judgment about standard coverage and measurement quality for each instrument relative to each standard. On the right, summarize what will be needed to adequately measure candidate performance for each standard.

Collaborate with teacher educators, arts and science faculty and school practitioner colleagues to assess each of your existing program exit performance measures and determine what additional work needs to be done to adequately address all teacher standards.

Table 2 shows a hypothetical example of WKU's review of its existing four performance measures and what the collaborative task force recommended as Western's assessment development needs.

The notes in the far right column of Table 2 summarize the existing status of performance assessments at Western relative to each of the nine Kentucky new teacher standards with the use of the four assessment instruments already adopted for all preparation programs. In this example, a review of standard coverage and instrument quality suggests several gaps in coverage and measurement quality. The most obvious gap in Western's present assessment package from this review is for *Standard III Collaboration* for which no assessment had been developed. Also, adequate coverage for *Standard IX Technology* in Western's Teacher Work Sample and Classroom Observation Record is questionable. While use of technology is addressed to some degree, not all expectations of the state's technology standard are being addressed.

With respect to the measurement quality of existing performance assessments instruments, both the Classroom Observation Record and the Portfolio Professional Development Plan have deficiencies and need work. The instrument quality of the Teacher Work Sample is adequate and that of the PRAXIS II is not questioned because of its published credibility data.

Thus, the review by the task force brings attention to the need for improving the Classroom Observation Record to include a well-developed rubric and the need to add more structure and a rubric to the existing professional development plan assigned task. In addition, it is evident that additional assessment instruments (performance tasks and rubrics) are needed for the collaboration and technology standards.

**Table 2. New Teacher Standards versus Assessment Measure Matrix (WKU Example)**

Kentucky Teacher Standards	Existing WKU Assessment Measures								Summary of Development and/or Improvement Needs
	Classroom Observation Record (COR)		Teacher Work Sample		Praxis II		PDP Plan Portfolio		
	Standard Coverage	Instrument Quality	Standard Coverage	Instrument Quality	Standard Coverage	Instrument Quality	Standard Coverage	Instrument Quality	
<b>Standard I</b> Knows Teaching Content	M	L	M	H	H	H	O	O NA	<ul style="list-style-type: none"> <li>Standard coverage: H</li> <li>Instrument quality: TWS &amp; Praxis II – H; COR needs a rubric</li> </ul>
<b>Standard II</b> Designs Instruction	M	L	H	H	O	NA	O	NA	<ul style="list-style-type: none"> <li>Standard coverage: H</li> <li>Instrument Quality: TWS – H COR needs a rubric</li> </ul>
<b>Standard III</b> Creates Learning Environment	H	L	L	H	O	NA	O	NA	<ul style="list-style-type: none"> <li>Standard coverage: H</li> <li>Instrument Quality: COR needs a rubric</li> </ul>
<b>Standard IV</b> Implements Instruction	H	L	H	H	L	H	O	NA	<ul style="list-style-type: none"> <li>Standard coverage: H</li> <li>Instrument Quality: TWS &amp; Praxis II – H; COR needs a rubric</li> </ul>
<b>Standard V</b> Assesses Learning	H	L	H	H	L	H	O	O	<ul style="list-style-type: none"> <li>Standard coverage: H</li> <li>Instrument Quality: TWS &amp; Praxis II – H; COR needs a rubric</li> </ul>
<b>Standard VI</b> Reflects and Evaluates Teaching and Learning	L	L	H	H	O	H	O	NA	<ul style="list-style-type: none"> <li>Standard coverage: H</li> <li>Instrument Quality: TWS – H; COR needs a rubric</li> </ul>
<b>Standard VII</b> Collaborates with Parents and Professionals	O	NA	O	NA	O	NA	O	NA	<ul style="list-style-type: none"> <li>Standard Coverage: O</li> <li>Instrument Quality: No instrument; a performance task and rubric need to be developed</li> </ul>
<b>Standard VIII</b> Engages in Professional Development	O	NA	L	H	O	NA	H	L	<ul style="list-style-type: none"> <li>Standard coverage: H</li> <li>Instrument Quality: TWS – H; PDP Plan task needs more structure and a rubric</li> </ul>
<b>Standard IX</b> Uses Technology for Instruction	M	L	M – L	H	O	N/A	O	NA	<ul style="list-style-type: none"> <li>Standard coverage L – M; needs more performance assessment opportunity for technology</li> <li>Instrument Quality: TWS – H; COR needs a rubric</li> </ul>

**Key:** H = High, M = Moderate, L = Low, O = Zero or Non-existent, N/A = Not Applicable

### Step 3. Create an Assessment Development Work Plan

Steps 1 and 2 helped you make sure that all standards are addressed with your institution's assessment package and that you have one or more assessments to adequately measure performance on each standard. The example of steps 1 and 2 at Western Kentucky University produced consensus about the need to improve the Classroom Observation Record and the Professional Development Plan as assessment instruments and to design additional "new" assessments for collaboration and use of technology.

Once you are satisfied that you have identified the complete set of instruments, you need to address all performance standards and you need to ask two questions that will help you plan the improvement of existing instruments and the design of new tasks, prompts and rubrics:

Question 1: Does the performance task and/or prompt provide the candidate a good opportunity to demonstrate the attainment of the standard? Note that what you ask candidates to do requires some definition and structure. Simply asking a candidate to teach a lesson without some specific parameters or instructions opens the door for the candidate to choose a lesson or classroom experience that may not allow or provide the opportunity to demonstrate the specific knowledge, skills, and dispositions called for in a teaching standard.

Question 2: Does the assessment enable raters, scorers and supervisors to judge consistently levels of performance for the standard? Performance tasks require good rubrics or clear instructions for judging levels of performance. Rating scales with labels that only say "below standards," "meets standards," and "exceeds standards" are too ambiguous to enable a rater to make consistent judgments. Good rubrics consist of specific behavioral indicators that provide evidence of meeting standards. Well defined rubrics that address each standard separately or allow the teasing apart of standards are essential for ascertaining that a performance demonstrates the meeting of a standard.

For any assessment, if you answered "no" to Question 1, then the problem lies with your task prompt; if you answered "no" to Question 2, then the problem lies with your rubric. If you answered "no" to both questions, then the performance assessment needs a full revision related to that standard.

Based on how you answered each question for each assessment, create a second matrix to show the development work needed to have quality assessment measures that address all teaching standards teacher candidates are required to meet in your program. For this matrix, in the left-hand column, list your existing performance assessment measures that need more work and the projected new assessment measures. Head columns with the standards that will be the focus of development and the types of development or improvement needed for each performance assessment. Table 3 presents a hypothetical example from Western Kentucky University. The matrix becomes the **work plan** that shows the development work that needs to be done on performance assessments so that data can be valid and reliable. This matrix will provide guidance for performance assessment development to meet identified needs.

**Table 3. Assessment Development Matrix for Performance Assessment Work Plan (WKU Example)**

<b>Assessment Measure to be Improved or Developed</b>	<b>Standards To Be Addressed</b>	<b>Needs Development From Scratch</b>	<b>Needs Full Revision</b>	<b>Needs Work on Task or Prompt</b>	<b>Needs Work on Rubric</b>
Classroom Observation Record	1, 2, 3, 4, 5, 6, 9				√
Professional Development Plan Task	8		√	√	√
Collaboration Task	7	√		√	√
Technology Task	9	√		√	√

**Key:** √ represents identified development needs. Numbers represent the need related to a particular teaching standard.

**Expected Product of this Component:**

A completed assessment development matrix that serves as a work plan to show what specific development work needs to be done to (a) improve existing assessment measures and (b) develop new assessment measures to address all teaching standards and provide consistent judgments about levels of performance on each standard.

**Tips and Good Advice for this Component:**

1. Review exemplars of quality performance assessments that have prompts and rubrics before you examine your existing set of assessments. (For example, to see the Renaissance Teacher Work Sample prompt and rubric, go to: <http://fp.uni.edu/itq/>).
2. Adopt a glossary of terms and discuss key concepts and terms among colleagues who will review existing assessments and suggest improvements. (For example, to see the glossary associated with the Renaissance Teacher Work Sample, go to: <http://fp.uni.edu/itq/>)
3. Be sure to have a mix of teacher educators, arts and science faculty and school practitioners to help you with this examination process. You need the voices of pedagogy, academic content, and classroom reality to develop good performance assessments.
4. Be sure to designate a facilitator for this component who has the skills and authority to engage different voices and move the process of this component to completion in three months or less.

## **Component 2: Designing Performance Assessments**

***Key Question: How do we create performance assessments that have high potential to measure fairly and consistently identified teaching standards?***

**Task for this Component:**

Design teacher candidate assessment measures that address specific teaching standards and have the ability to reliably identify levels of performance.

**Rationale for this Component:**

Standards-based teaching and learning demands the alignment of standards of teaching performance with assessments that measure progress towards standards and instruction that develops candidate performance towards standards. Once standards of performance have been identified, the development of quality assessment measures is key to translating broad outcomes into measurable skills, behaviors and products, identifying candidate progress towards teaching standards and guiding instruction that develops high levels of performance. Like a house designed to withstand the challenges of weather and time, performance assessments must have a solid foundation and contain quality components to withstand challenges to validity, bias, and consistency of judgments.

Building and establishing credibility of candidate performance measures are not only a requirement of the professional community accrediting agencies, they are essential for improving and producing high levels of teaching performance and P-12 learning.

As standards-based teaching and learning have been adopted in P-12 schools across the nation, valid and reliable measures of P-12 learning have received increased attention and have been given much weight in determining student progress and school quality. The same trend is true for standards-based teacher preparation. Our understanding of what knowledge, skills, and dispositions are necessary for teacher success in the classroom will be limited or enhanced depending on our progress in developing and using quality performance assessments.

Credible performance data are essential for both demonstrating program accountability and guiding program improvement. Such data can only be available if quality performance assessments are developed. Experience indicates that the upfront investment of resources to design quality assessments produces benefits that are well worth the time, effort and cost.

**Process for this Component:**

Creating quality assessments that directly address the teaching performance(s) identified by standards, that identify and differentiate between different levels of candidate performance, and that provide candidates the opportunity to demonstrate performances related to standards requires three steps:

- Identify the key indicators that will be used to judge performance on each standard.
- Develop rubrics that provide clear guidelines for ascertaining various levels of performance.
- Design or modify tasks and prompts related to standards, indicators and rubrics so that they provide teacher candidates the best opportunity to demonstrate the desired performance.

This component will guide you toward successful completion of these steps.

### **Step 1. Identify Key Indicators of Performance**

Once there is a complete set of exit performance standards all teacher candidates are to demonstrate at the exit from a teacher preparation program, obtain consensus among colleagues and program stakeholders about the factors or indicators that should be the focus of performance relative to the standard and “looked for” in judging performance. A key aspect related to this step is working with colleagues and stakeholders to define and describe terms embedded in the indicators to ensure a shared understanding. For example, below are three examples of sets of performance indicators from three teaching standards evaluated in the Renaissance Partnership Teacher Work Sample. For the first performance indicator below, “alignment with learning goals and instruction,” it would be important to define the term, “learning goals,” to help all stakeholders know what does/does not constitute a learning goal.

**Teaching Standard.** The teacher uses multiple assessment modes and approaches aligned with learning goals to assess student learning before, during, and after instruction.

Performance Indicators:

- Alignment with learning goals and instruction
- Clarity of criteria and standards for performance
- Multiple modes and approaches
- Technical soundness
- Adaptations based on the individual needs of students

**Teaching Standard.** The teacher designs instruction for specific learning goals, student characteristics and needs and learning contexts.

Performance Indicators:

- Alignment with learning goals
- Accurate representation of content
- Lesson and unit structure
- Use of a variety of instruction, activities, assignments and resources
- Use of contextual information and data to select appropriate and relevant activities, assignments and resources
- Use of technology

**Teaching Standard.** The teacher uses assessment data to profile student learning and communicate information about student progress and achievement.

Performance Indicators:

- Clarity and accuracy
- Alignment with learning goals
- Interpretation of data
- Evidence of impact on student learning

The above three teaching standards are assessed in the Renaissance Partnership institutions using a teacher work sample. The performance indicators are the basis for developing scoring rubrics described in the next step.

At Western Kentucky University, teacher candidates must also address teaching standards that cannot be directly addressed in teacher work samples. Below is an example of a teaching standard on classroom learning environment and another on collaboration:

**Teaching Standard.** The teacher creates and maintains a learning climate for students.

Performance Indicators:

- Communicates higher expectations
- Supports diversity and individual needs
- Uses positive classroom management techniques and strategies
- Facilitates mutual respect among students
- Employs creative and flexible use of time and materials
- Supports instruction through the creation of flexible and safe physical space

**Teaching Standard.** The teacher collaborates with colleagues, parents, and others to facilitate student learning.

Performance Indicators:

- Identifies situations when and where collaboration will enhance student learning
- Develops a plan for collaboration
- Facilitates collaborative activities
- Analyzes results of collaboration

The Renaissance Partnership has opted for seven process standards for teacher work samples and has identified a set of performance indicators for each standard. The standards and indicators can be downloaded by logging on to the Renaissance Partnership Project website at <http://fp.uni.edu/itg>. Similarly, Western Kentucky University has adopted the performance indicators for nine teaching standards required of new teachers. These can be viewed at the Kentucky Education Standards Board web site <http://kyepsb.net>.

## **Step 2. Develop Rubrics to Judge Candidate Performance**

Rubrics provide explicit instructions about judging levels of candidate performance. To be useful, rubrics must: (a) focus on the teaching standard and thus address the identified performance indicators; (b) distinguish between different levels of performance (describe what would likely be observed or produced at different stages of development toward “proficient” performance on the standard); and (c) use terms and descriptions that have a common meaning to teacher candidates, teacher educators, arts and science faculty and school practitioners. Experience has shown that a shared glossary of terms is always beneficial.

Writing good rubrics is very difficult but important. While most of the time rubrics are developed with specific teaching tasks or situations in mind (e.g., teacher work samples, classroom observations, stand-alone teaching tasks), it is recommended that the first draft of a rubric be developed by simply focusing on descriptions of different levels of performance.

While the number of performance levels in a rubric is arbitrary, a large proportion of the professional community, including the National Board of Professional Teacher Standards, utilizes four. Table 4 presents examples of the labels used to describe teaching performance levels by several different groups and teacher preparation institutions. While different terms are used to describe these four levels, the concept or idea behind these four different levels is very similar.

**Table 4. Labels Used by Different Groups for Four Levels of Teaching Performances**

Organization	Level 1	Level 2	Level 3	Level 4
Renaissance Partnership	Beginning	Developing	Proficient	Expert
Western Kentucky	Standard not demonstrated	Standard partially demonstrated	Standard demonstrated	Standard exceeded
National Board*	1	2	3	4
Idaho State University	Beginning	Developing	Proficient	Exemplary

\*2.75 is considered passing or “proficient”

For each of the standards you have identified and/or are required to address in your teacher preparation program, develop or borrow descriptions of performance related to the different levels you choose to adopt at your institution or organization. Tables 5 and 6 provide two examples of fully developed rubrics.

**Step 3. Design/Modify Tasks and Prompts Related to Standards, Indicators and Rubrics**

The third step in designing valid and reliable performance assessments is to structure a task, activity or situation that will provide an opportunity for the teacher candidate to demonstrate the performances required by the teaching standard.

In the past, teacher educators have required teacher candidates and interns to produce general portfolio exhibits as evidence of teaching performances. They also have made classroom visits and recorded their observations without any common understanding with the teacher candidates about the performances (skills, behaviors, interactions) they were looking for relative to teaching standards. Whereas the open ended nature of portfolio exhibits and classroom demonstrations of instruction offer maximum flexibility to the candidate, it gives little consideration to a situation or a task that provides the candidate the best opportunity to demonstrate the specific performances required by a teaching standard.

Providing some structure to a teaching situation or task focuses the candidate on activities and products that more clearly demonstrate levels of performance directly related to standards and indicators. In the Renaissance Teacher Work Sample a series of seven teaching tasks are given to the candidate (one for each teaching process standard) that includes planning, teaching, analyzing student learning, and evaluation of the results of a unit of instruction. Western Kentucky University uses the Renaissance teaching tasks associated with the teacher work sample as assessments of performance for some Kentucky standards but, in addition, structures three lesson demonstrations, and requires stand-alone performance tasks for standards on professional development, collaboration and use of technology.

**Table 5. Example Rubric – TWS Assessment Plan Rubric**

<b>TWS Standard</b>				
<i>The teacher uses multiple assessment modes and approaches aligned with learning goals to assess student learning before, during, and after instruction.</i>				
<b>Rating → Indicator ↓</b>	<b>1 Indicator Not Met</b>	<b>2 Indicator Partially Met</b>	<b>3 Indicator Met</b>	<b>Score</b>
<b>Alignment with Learning Goals and Instruction</b>	Content and methods of assessment lack congruence with learning goals or lack cognitive complexity.	Some of the learning goals are assessed through the assessment plan, but many are not congruent with learning goals in content and cognitive complexity.	Each of the learning goals is assessed through the assessment plan; assessments are congruent with the learning goals in content and cognitive complexity.	
<b>Clarity of Criteria and Standards for Performance</b>	The assessments contain no clear criteria for measuring student performance relative to the learning goals.	Assessment criteria have been developed, but they are not clear or are not explicitly linked to the learning goals.	Assessment criteria are clear and are explicitly linked to the learning goals.	
<b>Multiple Modes and Approaches</b>	The assessment plan includes only one assessment mode and does not assess students before, during, and after instruction.	The assessment plan includes multiple modes but all are either pencil/paper based (i.e. they are not performance assessments) and/or do not require the integration of knowledge, skills and reasoning ability.	The assessment plan includes multiple assessment modes (including performance assessments, lab reports, research projects, etc.) and assesses student performance throughout the instructional sequence.	
<b>Technical Soundness</b>	Assessments are not valid; scoring procedures are absent or inaccurate; items or prompts are poorly written; directions and procedures are confusing to students.	Assessments appear to have some validity. Some scoring procedures are explained; some items or prompts are clearly written; some directions and procedures are clear to students.	Assessments appear to be valid; scoring procedures are explained; most items or prompts are clearly written; directions and procedures are clear to students.	
<b>Adaptations Based on the Individual Needs of Students</b>	Teacher does not adapt assessments to meet the individual needs of students or these assessments are inappropriate.	Teacher makes adaptations to assessments that are appropriate to meet the individual needs of some students.	Teacher makes adaptations to assessments that are appropriate to meet the individual needs of most students.	

(Reproduced from The Renaissance Partnership For Improving Teacher Quality Teacher Work Sample: Performance Prompt, Teaching Process Standards, Scoring Rubrics)

**Table 6. Example Rubric – Pathwise “The Teacher Creates and Maintains a Learning Climate for Students” Rubric**

<b>Standard Addressed</b> <i>The teacher creates and maintains a learning climate for students.</i>			
<b>Rating→ Benchmark ↓</b>	<b>3 Benchmark demonstrated</b>	<b>2 Benchmark partially demonstrated</b>	<b>1 Benchmark not Demonstrated</b>
<b>lia. Communicates high expectations for all students</b>	Teacher sets significant and challenging goals for students and then verbally/nonverbally communicates confidence in students' ability to achieve these goals.	Teacher verbally/nonverbally communicates confidence in students' ability to achieve lesson goals; however, goals lack significance and challenge.	Teacher fails to verbally/nonverbally communicate confidence in students' ability to achieve lesson goals or verbally/nonverbally communicates confidence in some students' ability and limited/no confidence in others.
<b>lib. Supports student diversity and addresses individual needs</b>	Teacher consistently demonstrates a sensitivity to student diversity and individual needs in a variety of ways such as nonverbal, verbal, and written communication, grouping practices, selection of activities and materials, and room arrangement.	Teacher demonstrates sensitivity to some areas of student diversity and individual needs but fails to respond to others.	Teacher seldom, if ever, demonstrates sensitivity to student diversity and individual needs.
<b>lic. Uses positive classroom management techniques that foster self-control and self-discipline to create and sustain a climate that motivates students to learn</b>	Classroom climate is conducive to teaching and learning. Most students exhibit age-appropriate levels of self-control and self-discipline. Teacher deals with lapses in appropriate behavior in a positive fashion.	Classroom climate is usually conducive to teaching and learning. Some students exhibit frequent lapses in age appropriate self-control and self-discipline. Teacher usually responds to lapses in appropriate behavior in a positive fashion.	Classroom climate is not conducive to teaching or learning. Many students fail to exhibit age-appropriate levels of self-control and self-discipline. Teacher ignores misconduct or responds in punitive/threatening fashion.
<b>lid. Facilities mutual respect among class members through cooperative and independent learning activities</b>	Teacher uses cooperative and independent activities to create opportunities for student interactions and consistently monitors these interactions to reinforce appropriate behaviors and discourage inappropriate behaviors.	Teacher uses cooperative and independent activities to create opportunities for student interaction however monitoring of these interactions is inconsistent, allowing inappropriate behaviors to often occur unchallenged.	Teacher provides limited opportunities for student interactions AND/OR fails to monitor interactions to ensure appropriate behavior.
<b>lie. Employs creative and flexible use of instructional time and materials</b>	Teacher effectively varies the use of time and materials in response to student learning needs and to facilitate instruction.	Teacher makes some variation in the use of time and materials to facilitate instruction and respond to student learning needs.	Teacher makes little or no variation in the use of time or materials.
<b>lif. Supports instruction through the creative, flexible, and safe use of physical space</b>	Teacher has done an outstanding job of using the available physical space in ways that facilitate instruction and create a pleasant place for student learning.	Teacher has made an effort to use the available physical space in ways that contribute to instruction and student learning.	Teacher has made little or no effort to use available physical space in ways that contribute to instruction or student learning.

(Reproduced from Pathwise by Charlotte Danielson, Educational Testing Service)

Measuring candidate progress includes other assessment components such as structural interviews and journal entries to assess performances related to teaching standards. These can be configured in different formats and venues but the important concepts to remember are that all performance tasks, portfolio entries or observed teaching situations need structure to ensure the candidate has the opportunity to demonstrate the required and expected performances.

For all the standards you identified in your work plan matrix of the previous component, identify the performance task or set of tasks that will provide the best opportunity for a teacher candidate to demonstrate performance related to the standard or standards. Pay particular attention to the performance indicators you have identified for each standard and think of what activity, situation, or assignment (task) you could design in which a teacher candidate can provide an authentic demonstration of the desired performance. The performance task statement should be a succinct statement that describes an activity or assignment and the general conditions under which the activity or assignment is to be carried out. The task statement should be detailed enough to ensure an excellent opportunity to demonstrate performance but general enough to provide teacher candidates' flexibility to show their own creativity in performance.

Following the performance task statement, provide the candidate with a set of suggestions or directions for optimum demonstration of the desired performance(s). Prompts may be directions, questions to think about or just good advice on how to exhibit high levels of performance. The length and extent of the prompts that should follow a performance task statement will depend on the complexity of the task and the developmental level of the teacher candidate. Remember, the purpose of both the task and prompts that follow is to maximize the opportunity for the candidate to demonstrate performance in an efficient manner.

On the next pages are examples of performance tasks and prompts for two of the teaching standards and scoring rubrics shown earlier. The first (designing a unit of instruction) is a component of the Renaissance Teacher Work Sample package. The second is a stand-alone task and prompt created by colleagues at Western Kentucky University to address the state standards on collaboration.

**Expected Products from this Component:**

A set of performance assessment measures that can be used with teacher candidates to assess and judge levels of candidate performance with respect to all program standards of teaching performance. Each performance measure includes a

- statement about which standards are being addressed by the assessment task, measure or instrument;
- performance task and prompt or a description of the conditions/situation under which the performance is to be demonstrated or exhibited and suggestions that encourage an authentic demonstration of the desired performance; and
- scoring rubric that identifies key factors or indicators of performance to observe or “look for” in the performance exhibit followed by a description of performance at different levels of proficiency and/or instruction about how to judge candidate performance on the specified task(s).

## Design for Instruction

### TWS Standard

*The teacher designs instruction for specific learning goals, student characteristics and needs, and learning contexts.*

### Task

Describe how you will design your unit instruction related to unit goals, students' characteristics and needs, and the specific learning context.

### Prompt

- **Results of pre-assessment.** After administering the pre-assessment, analyze student performance *relative to the learning goals*. Depict the results of the pre-assessment in a format that allows you to find patterns of student performance relative to each learning goal. You may use a table, graph, or chart. Describe the pattern you find that will guide your instruction or modification of the learning goals.
- **Unit overview.** Provide an overview of your unit. Use a visual organizer such as a block plan or outline to make your unit plan clear. Include the topic or activity you are planning for each day/period. Also, indicate the goal or goals (coded from your Learning Goals section) that you are addressing in each activity. Make sure that every goal is addressed by at least one activity and that every activity relates to at least one goal.
- **Activities.** Describe at least three unit activities that reflect a variety of instructional strategies/techniques and explain why you are planning those specific activities. In your explanation for each activity, include:
  - how the content relates to your instructional goal(s),
  - how the activity stems from your pre-assessment information and contextual factors,
  - what materials/technology you will need to implement the activity, and
  - how you plan to assess student learning during and/or following the activity (i.e., formative assessment).
- **Technology.** Describe how you will use technology in your planning and/or instruction. If you do not plan to use any form of technology, provide your clear rationale for its omission.

**Suggested Page Length:** 3 + visual organizer

## **Intern Performance Assessment Component: Collaboration**

### **Standard Addressed**

*VI. The teacher collaborates with colleagues, parents and others.*

### **Task**

Collaborate with parents, guardians, or primary caregivers and one or more other professionals to design and implement a special learning plan for two different students with special needs. Report on your process and the impact of each plan on student learning.

### **Prompt**

- Using your Contextual Factors data and other sources, identify students in your class(es) with identified special needs whose learning would be enhanced by collaborative efforts. These students could be students who have special needs because of learning challenges (e.g., students with IEPs or 504 plans or ESL students) or students whose special needs are a result of his/her strengths (e.g. GSSP students). From the identified special needs students, select two students who will be the focus of your collaborative efforts.
- For each of the students you selected:
  - Initiate a collaborative effort involving the student, one or more parents, guardians or primary caregivers and at least one other professional to design and implement a plan of special strategies and activities designed to impact student learning. For this collaborative plan to impact student learning, it should address a significant need and be of sufficient duration for measurable impact on learning to occur (six to twelve weeks).
  - Be sure to collect and use student data to both select/design the plan's strategies and activities and assess student learning progress.
  - Conduct an initial collaborative planning meeting, at least one interim progress check meeting, and a final assessment and reflection meeting.
- Your Collaboration Exhibit should be included in your Cycle 3 Exit Portfolio. This exhibit should include:
  - A brief description of your two collaborative efforts: the objectives of each plan, the participants in each collaboration and why you selected each of the students involved.
  - A brief description of the collaboratively developed "learning" plan for each of the two students.
  - A brief description of each interim planning meeting and the final meeting.
  - The results of the collaborative plan on each student's learning.
  - An analysis and reflection on the collaborative efforts that use student data to determine the impact of efforts and identify possible next steps.
- A brief progress report on your collaborative efforts should be presented in your Entry and Mid-Point Portfolios to inform your committee of your progress in addressing this standard.

**Tips or Good Advice in Developing this Component:**

1. Make the performance tasks, prompt, and rubrics as succinct as possible. Long and complex assessment instruments may confuse candidates and faculty.
2. Use terms that everyone understands and ensure alignment of terms between standards, tasks, prompts, and rubrics. Again, a glossary of terms is highly recommended.
3. Give a copy of your first drafts of tasks, prompts and rubrics to a few teacher educators, school practitioners and teacher candidates who have not been involved in the development and get their input and feedback for clarity, authenticity and alignment with the standard(s).
4. Be prepared for frustrations and challenges. Designing and developing good performance assessments challenges even the brightest among us. If the process seems easy, you probably do not fully comprehend the purpose and function of this component in the performance assessment process.
5. Be prepared to make minor adjustments to the rubrics as you begin using them. Developing a good rubric is iterative—you aren't likely to "get it right" the first time. However, there may come a time to "stand firm" with colleagues and other stakeholders on making further changes until a rubric has had a sufficient opportunity to work.

## **Component 3: Initial Production of Candidate Performances Analysis**

**Key Question:** *What are the essentials to collect initial candidate performances in preparation for credibility studies?*

**Task for this component:**

Provide drafts of assessment instruments, supporting materials and specific directions to a set of typical teacher candidates and collect a set of completed performances that represent typical performance behavior.

**Rationale for this component:**

Performance assessments frequently require extensive guidelines and directions for completion. All teacher candidates must be given a copy of the guidelines for the performance assessment delineating the required tasks and the necessary steps for preparing them. For the performance assessment to be a valid measure of candidates' abilities to meet the targeted standards, it is essential the requirements be clearly communicated to all candidates completing the performance assessment. In addition, the faculty members who mentor teacher candidates and members of the professional community who may be selected to judge candidate performance levels all must hold similar conceptions of the task requirements and performance indicators (McConney & Ayres, 1998). Hence, acquainting all members of the professional community with the performance assessment is a critical initial undertaking.

The performances can be collected from all teacher education candidates for whom the assessment is intended or from a representative sample. A representative collection of performance is important for determining inter-rater agreement and to have examples for the development of credibility evidence. Larger institutions may be able to collect a large enough sample (at least 50) of candidate performances in a single semester. Smaller units may need to collect performances from all of their candidates for an entire year or more to have a sufficient numbers for a credibility evidence study. Determine whether all candidates for whom the assessment is intended will complete the assessment or only a representative sample. Pilot studies may be conducted with non-representative groups of candidates (such as a few sections of a course) or with only candidates enrolled in specific programs (such as elementary education or special education). However, the participants in conclusive studies must be representative of all candidates for whom the performance assessment is intended.

To reduce bias in scoring, code numbers should be the only identifying information on the performances. This helps to insure fairness by removing potential biasing factors and other extraneous information. This also protects the confidentiality of candidate performances.

**Process for this Component:**

**Step 1. Provide Candidates with Performance Assessment Prompts and Rubrics**

To aid clear communication of the performance tasks, provide all teacher candidates a copy of the performance assessment guidelines (For an example set of guidelines for the Renaissance Teacher Work Sample, visit the Renaissance Partnership Project web site at <http://fp.uni.edu/itq>). The guidelines must specify the tasks to be performed and the expected documentation. Be sure to share copies with faculty members who mentor candidates as well

(See also the rater training component presented later in this manual.). Furthermore, provide teacher candidates with a complete copy of the scoring rubric that will be used to assess their performance. The candidates need to be aware of all aspects of how their performances will be judged.

### **Step 2. Initiate Candidate Performance Assessments**

Instruct your candidates to complete the required performance assessments. Determine which teacher candidates will complete the performance assessment and at what point during their program. All aspects of the circumstances of each performance must be determined. Some performance assessments can be completed as part of course requirements; in which case, the course instructors will tell the candidates about when and how to complete the assessment. Other performance assessments, such as teaching observations, can only be carried out after appropriate field placements have been arranged and supervisors have been assigned. It is a good idea to develop a program handbook that explains your assessment system to your candidates. It is also a good idea to specify all required assessments in your official institutional catalog.

Be sure to provide candidates the appropriate observation and evaluation conditions to complete the required performance assessment or engage in the required performance. Also, allot candidates sufficient time to complete each performance and prepare documentation. The due date and procedures for returning the completed assessment must also be specified.

### **Step 3. Mentor Candidates**

Because performance assessments are complex, determine a process for mentoring candidates. Course instructors and cooperating teachers who mentor candidates must also be trained, so they can provide relevant assistance. Clear expectations must be set forth for the amount and kind of assistance candidates are allowed when completing the performance assessment. Determination must be made of the amount of independence expected when performances are to be used as demonstration of candidates' abilities to meet standards. Mentors should be available to answer questions, but should not assist the candidates in completing the tasks beyond the level specified for that assessment.

### **Step 4. Develop Performance Collection and Evaluation Preparation Mechanisms**

Be sure to develop a mechanism to collect candidate performances. Possible means of collection include having students submit two copies of their performance—one for classroom grading that will be returned to the student and the other for institutional collection of credibility evidence. Another approach would be for faculty members to select a subset of performances from their classes to be copied and submitted for future evaluation studies.

For conducting credibility evidence studies, it is important that all identifying information be removed from the performances. Candidate information can be entered into a separate database linked to coded identifiers—usually numbers assigned to each candidate.

### **Expected Product of this Component:**

A representative set of performance exhibits that can be scored and benchmarked or classified with respect to “levels” of performance.

### **Tips and Good Advice for this Component:**

1. It is a good idea to develop a program handbook that explains your assessment system to your candidates. It is also a good idea to specify all required assessments in your official institutional catalogue.

2. Guidelines and scoring rubrics for all performance assessments can be posted on your website using PDF files for the teacher candidates and faculty mentors to download.
3. Clarity can be improved by having faculty members who mentor teacher candidates keep track of frequently misunderstood directions and frequently asked questions. This information can be used to make improvements to the guidelines given to future candidates. Answers to frequently asked questions can also be posted on your website.
4. The development of an Assessment Committee can assist in formulating policies and procedures for administering assessments. Bi-monthly meetings to discuss needs, challenges, and successes should be considered. The committee can also develop an assessment system handbook and other materials for advising program candidates.
5. For complex performances completed independently, have candidates complete an affidavit attesting to the fact that the work presented is their own. Otherwise, for observable performances, proctors can be used while candidates complete the assessments or judges can view the performances directly in appropriate settings.
6. Performances should be collected at the end of each semester.
7. Use FileMaker Pro or another database software to generate a database containing candidate information and code identifiers for each performance. Enter as much information as possible into the database before the study begins.

## **Component 4: Planning Credibility Evidence Studies**

**Key Question:** *What steps are necessary for planning and conducting quality studies of credibility evidence?*

**Task for this Component:**

Design a work plan of activities, role responsibilities and timelines for conducting benchmarking or performance exhibits, validity studies and scorer agreement (generalizability) studies.

**Rationale for this Component:**

Development of an assessment system and quality performance assessments requires a culture of evidence. All assessments, including performance assessments, must meet technical standards, if they are to be used to make consequential decisions about candidate performance levels (American Educational Research Association, 1999; NCATE, 2000) and become accepted as measures of institutional and state standards. The goals for evidence gathering are to: (1) support the validity of the scoring rubric for the purpose of documenting candidates' abilities to meet program and state teaching standards targeted by the assessment; (2) demonstrate the assessment differentiates performance levels and to establish benchmarks of those levels along a continuum from beginning to highly accomplished performance; (3) determine whether the performance assessment can be feasibly and equitably administered, and (4) determine whether the performance assessment can be scored with sufficient inter-rater agreement to warrant its use in high-stakes decisions about the effectiveness of candidates' performance with respect to the targeted standards. In addition, support should be sought for the predictive validity of each performance assessment through follow-up studies of program graduates. Study plans and data collection can be organized according to local constraints and participant availability. The person(s) responsible for establishing credibility evidence for performance assessments must have a plan for collection of the performances, security of the performances, and the generation and storage of all materials. There should be a pre-determined database structure for each performance assessment for recording all data and participant demographic information. Finally, a plan should be determined in advance for the analysis of all data as well.

**Process for this Component:**

**Step 1. Appoint an Assessment Coordinator and Credibility Evidence Team**

Appoint an overall Assessment Coordinator and establish a credibility evidence team for each assessment. The teams can be composed of interested faculty members who have been involved with developing the assessment. The Assessment Coordinator should be a member of all teams and help to coordinate the separate studies across your multiple assessments.

Assign the responsibility to the Assessment Coordinator or other team leaders for obtaining approval from your institutional review board overseeing the use of human participants in research studies after the plans for your credibility studies have been written.

## Step 2. Design the Credibility Study

Include in the plans for credibility studies the elements common to the type of research study being conducted. Typical essentials include designation of the participants, the instruments, the data collection procedures, and the types of data analysis to be performed. Suggestions for each of these will be addressed in the components of this manual that follow.

Be sure to designate roles for all investigators involved in the study. Determine who will lead the process; who is responsible for insuring all necessary materials, equipment, and supplies are available on site; who will notify participants of the date, time and location for the study; who will enter the data into a database or statistical package as it is collected; who will analyze the data afterwards; and who will arrange refreshments, etc.

## Step 3. Develop a Data Collection Timeline and Outline

Table 7 presents an example outline for day one of a benchmarking and credibility evidence study of the Idaho State University Teacher Work Sample (Denner, Salzman & Bangert, 2001). Table 8 presents an outline for day two of the study.

**Table 7. Example Outline for Day One of a Benchmarking and Credibility Study**

Time	Outline for Day One
8:30 a.m.	Welcome and Introductions
8:45 a.m.	Participant Demographic Questionnaire
9:00 a.m.	Purpose of benchmarking and overview of the Teacher Work Sample assessment (Examination of the standards, guidelines, and scoring rubrics).
10:00 a.m.	BREAK
10:15 a.m.	Evaluator Guidelines & Anti-Bias Training Prepare for Benchmarking: <ul style="list-style-type: none"><li>• Read holistic rubric and highlight dimensions of four score categories.</li><li>• Make group assignments.</li></ul>
11:00 a.m.	Benchmarking Part I - Quick Read: <ul style="list-style-type: none"><li>• Work in groups of three raters to “quick read” assigned teacher work samples.</li><li>• Discuss categorization with group members.</li><li>• Group must reach consensus regarding score category placement.</li><li>• Place the work sample in the correct category pile on marked tables.</li></ul>
12:00 p.m.	LUNCH
1:00 p.m.	Benchmarking Part I Continues Until Completed.
2:00 p.m.	Benchmarking Part II - Selection of Exemplars <ul style="list-style-type: none"><li>• Work in groups of three raters (composition should be different from the morning session).</li><li>• Each group examines all work samples in a single category and selects five or exemplars of that category. Selection is by consensus.</li></ul>
4:00 p.m.	Raters individually complete a validity questionnaire.
4:30 p.m.	Closure

**Table 8. Outline for Day Two of a Benchmarking and Credibility Study**

Time	Outline for Day Two
Prior Evening	Investigators organize benchmarked performances into sets of 10 to 20 by randomly selecting performances from within each developmental level. In this way, representative sets of performances are organized for raters to score.
8:30 a.m.	Review of the scoring guidelines and the full scoring rubric. Assign raters to score benchmarked sets using the analytic scoring rubric.
9:00 a.m.	Analytic Scoring Three groups of 5 raters score sets of teacher work samples using the analytic scoring rubric. Each rater scores individually. Each group scores a different set of 10 teacher work samples (two beginning, three developing, three proficient, and two exemplary). The raters return completed scoring rubrics to investigators for data entry into a database located on a portable microcomputer.
12:00 p.m.	LUNCH
1:00 p.m.	Scoring Continues Until Finished
4:30 p.m.	Closure

**Expected Product of this Component:**

A completed work plan that identifies responsible persons for different roles and realistic dates and timelines for each of three credibility study processes: benchmarking of performances, determining validity of instrument with respect to standards and determining the ability of the instruments and scorers to produce consistent evaluation of performance exhibits.

**Tips and Good Advice for this Component:**

1. Have a graduate assistant available for miscellaneous activities and errands.
2. Consider providing lunch and refreshments during other breaks.

## **Component 5: Recruiting Qualified Raters**

**Key Question:** *What important characteristics should be considered to identify and recruit raters?*

**Task for this component:**

Recruit and identify a cadre of professionals who have a high potential to become qualified raters (scorers) and credible judges of performance exhibits.

**Rationale for this component:**

Performance assessment requires professional judgment; therefore, the judges must be credible experts. Criteria should be developed for determining rater qualifications. For example, teaching experience is a valuable background for scoring teacher work samples. Separate qualification criteria can be set for raters participating on scoring panels and raters recruited to evaluate the validity of the performance assessment. Rater experience mentoring candidates and scoring performances is likely to be an important factor (Dunbar, Koretz & Hoover, 1991). People who work closely with candidates make better raters because of their experience with the process. However, using outside raters adds credibility to the ratings and helps to gain wider support. Panels can be composed of mixed representatives such as cooperating teachers, program faculty members, and university supervisors.

Because performance assessments generally require multiple raters to achieve sound judgment, it will be necessary to recruit and train a large group of qualified raters. In addition, if a large number of performances are to be assessed, then multiple panels of qualified raters will be needed for efficient and timely assessment.

**Process for this component:**

**Step 1. Gain Institutional/Program Commitment**

Acquaint the professional community with the performance assessment. Broad awareness and commitment to your performance assessment system will aid your efforts to recruit qualified raters.

**Step 2. Develop Rater Qualification Criteria**

Develop criteria for determining rater qualifications. The qualification may vary for each performance assessment. The Assessment Coordinator or Assessment Committee should lead the process of establishing the appropriate criteria for each assessment.

Based on the criteria, send invitations to potential raters explaining the criteria and soliciting their participation in the study. Raters must be willing to commit a substantial amount of time to the scoring process so be sure to inform them of this time commitment in the invitation.

**Step 3. Include Multiple Constituents as Raters**

Consider including representatives from the following groups as raters in your initial credibility studies:

1. course instructors who mentor candidates,
2. university supervisors who evaluate candidate performance during student teaching internships,
3. cooperating teachers who work with program candidates,
4. public school administrators, and

5. university faculty members from other colleges who teach program candidates in content specialty areas.

Panel members should also be chosen to maintain gender, ethnic, racial, and teaching-setting diversity, if possible.

#### **Step 4. Collect Demographic Information**

Collect and report demographic information about the panel members (Crocker, 1997).

#### **Expected Product of this Component:**

A roster of professionals with the following characteristics: (a) The professionals represent key professional groups; (b) The professionals have a high potential to become expert and qualified raters; (c) The professionals have agreed to the conditions and responsibilities of the scoring project; and (d) The roster has a sufficient number of participants to provide credible scoring results.

#### **Tips and Good Advice for this Component:**

1. Hand select your first raters from your “best” colleagues—conscientious, cooperative, high frustration tolerance—with the goal of eliminating any factors that might compromise the success of your first attempt to collect credibility data.
2. Set the date for your credibility studies well in advance to ensure rater availability.
3. Offer an incentive for participation. If monies are available, consider paying the raters for their services. Faculty members can be given workload credit to include in their documentation for annual evaluations and promotion and tenure portfolios. Public school teachers may need substitutes hired for them.
4. Ask the Dean or Assistant Dean to speak to the raters about the importance of the contributions they are making to the work of the college.

## **Component 6: Training Raters to Score Performances**

***Key Question: What qualities should a training and scoring session possess to increase the likelihood of collecting reliable and valid scoring data?***

### **Task for this component:**

Prepare raters to score performances with an initial four to six hours of training. This includes a review of general guidelines for scoring, consideration of potential biases in scoring, discussion of the scoring rubric and the meaning of terms and indicators, and most important, practice scoring the performances using the assessment rubric.

### **Rationale for this component:**

Because performance assessments require the judgments of expert raters to score the assessments accurately, rater inexperience and rater bias are potential sources of score invalidity and poor reliability for all performance assessments. Rater training will enhance understanding of the scoring process, scoring accuracy, and inter-rater agreement (Dunbar, Koretz & Hoover, 1991). Although potential raters can be screened for blatant bias and conflicts of interest affecting fairness, all raters will always bring their personal preferences and idiosyncrasies to the scoring task; therefore, it is also important to include anti-bias training as part of the training process. Review with scorers a general set of guidelines for scoring. This will aid them to view performances consistently. Finally, it is important for the raters to have an opportunity to practice scoring performances using the assessment rubric. The following section describes a training process devised and used by Denner, Salzman, & Bangert (2001).

### **Process for this component:**

#### **Step 1. Schedule the Training and Scoring Session**

Discuss and arrange with potential raters a schedule for training and scoring performances. The Assessment Coordinator (or team leader) should assemble the raters, materials, and trainers on the planned dates in accordance with a previously developed plan (See Component 4).

#### **Step 2. Help Raters Connect Standards, Prompts, and Rubrics**

During training, help raters understand clearly the relationships among the performance expectations of the standards, the directions, and prompts given to the candidates, and the ways in which the rubric captures each dimension of performances that meet the standards. The training should allow for discussion of scale values for each indicator and for the clarification of key terms. It should also include a demonstration of the expected procedures for recording scores and other pertinent information (such as rater identification codes) on the scoring rubric. If benchmark performances are available, they can be used to illustrate scale values for each indicator dimension and for practice scoring (The importance of benchmarking will be discussed in the next component).

#### **Step 3. Create a Roadmap of Evidence**

For complex performances, create a template or “Roadmap of Evidence” similar to the one shown below in Table 9 for the Renaissance Teacher Work Sample that shows raters where to look for the evidence related to each standard.

**Table 9. Renaissance Teacher Work Sample Roadmap for Locating Evidence**

TWS Sections→ Teaching Process Standards ↓	Contextual Factors	Learning Goals	Assessment Plan	Design for Instruction	Instructional Decision- Making	Analysis of Learning Results	Evaluation and Self- Reflection
<i>The teacher uses information about the learning-teaching context and student individual differences to set learning goals and plan instruction and assessment.</i>	X	X	X	X	X		
<i>The teacher sets significant, challenging, varied, and appropriate learning goals.</i>		X	X	X	X		
<i>The teacher uses multiple assessment modes and approaches aligned with learning goals to assess student learning before, during, and after instruction.</i>			X	X			
<i>The teacher designs instruction for specific learning goals, student characteristics and needs, and learning contexts.</i>				X			
<i>The teacher uses ongoing analysis of student learning to make instructional decisions.</i>				X	X		
<i>The teacher uses assessment data to profile student learning and communicate information about student progress and achievement.</i>						X	X
<i>The teacher reflects on his or her instruction and student learning in order to improve teaching practice.</i>						X	X

\*Large X indicates primary source of evidence; small x indicates other potential sources; blanks indicate areas that are NOT used for evidence.

#### **Step 4. Conduct Rater Anti-bias Training**

Engage the raters in anti-bias training. In this training, the raters are asked to list characteristics of excellent performances like those targeted by this assessment and characteristics of very poor performances. After these lists are completed, the raters are asked to discuss them with each other as a whole group or in smaller clusters. If any raters recognize a preferential characteristic after hearing another rater talk about it, they can add it to their own list. Raters are then asked to compare the characteristics they wrote on their personal lists to the standards targeted by the performance assessment. Any preferential characteristics not appearing in the standards (or scoring rubric indicators) are recorded on a “*Hit List of Personal Biases*” (Denner,

Salzman & Bangert, 2001, p. 293). It is important to remind the raters “to focus on the standards as the sole lens for scoring” the performance (p. 293). Instruct the raters keep their *Hit List of Personal Biases* available while scoring as a constant reminder to focus on the standards and indicators only.

### **Step 5. Review General Assessor Guidelines**

Review a prepared set of general assessor guidelines (Denner, Salzman & Bangert, 2001). The guidelines should address the following issues to help the raters to maintain a proper attitude toward performances while scoring:

- Respect for and confidentiality of all performances
- Recognition that quality teaching has many faces
- Security of the Performances
- Subtleties of scoring, such as, Halo and Pitchfork Effects, or being consistently too lenient, stringent, or tending toward the center.
- A reminder that the standards (rubric indicators) are the only lens for judging the performance.

### **Step 6. Rate and Discuss Common Performances**

Have the raters score one or more performances. Allow sufficient time for individual scoring. Ask the raters to discuss their ratings. The trainer(s) should also lead a whole group discussion. In addition to providing an opportunity for the raters to calibrate their ratings by learning from one another, these discussions can uncover unanticipated scoring difficulties. Agreements resolving these difficulties can be codified in a scoring guide for future raters or in some cases lingering disagreements may point to necessary revisions of the rubric or other aspects of the performance assessment. If benchmarked examples are available (see next Component), have the raters compare their scores to the benchmark scores.

### **Expected Product of this Component:**

Raters are trained and prepared to score performances.

### **Tips and Good Advice for this Component:**

1. It is important to use standard terms to refer to the various forms and performance tasks (Crocker, 1997).
2. Training for mentors can also include discussion of how to give candidates feedback regarding performance for persons charged with mentoring candidates.
3. It is important that the mentors understand the meaning of all performance indicators included in the scoring rubric and the descriptions of different performance levels.
4. Mentors can be trained to score the performance assessment and at the same time help to obtain credibility evidence for its technical merits.
5. Any raters who do not think they can bracket their bias should be excluded from panels used to establish credibility evidence and from any consequential decisions regarding candidate performance. These individuals could still be allowed to score as a means to learn how their scores compare to other raters, but their ratings should not be used to make decisions about candidate performance unless and until they demonstrate the ability to score without significant bias. Of course, the evidence may also reveal that other raters should be excluded in the future as well due to low scoring reliability.

## Component 7: Benchmarking Performances

**Key Question:** *What are the essential steps to identify and provide good exemplars of levels of performance – “benchmarks” – that guide candidate development and the assessment process?*

**Task for this component:**

Given a set of representative performance exhibits, identify good examples of performance at each of the levels defined by the rubric.

**Rationale for this component:**

Benchmarks are examples of actual performances that represent different proficiency levels against which candidates' performances can be judged. Because candidates go through different stages as they develop the knowledge and skills required by state and institutional teaching standards, benchmarks can be identified that exemplify those stages along a continuum of performance quality. Selected benchmark performances illustrate cut-off points between performance levels or more often exemplify typical performances near the middle of a performance level. Benchmark examples are useful for training and calibrating the judgments of performance assessors. The process of benchmarking itself can help to acquaint novice raters with the range of quality of the performances they will be expected to rate. Benchmark examples at higher performance levels are also useful as aspiration models for future candidates and as teaching illustrations for faculty members who mentor your candidates.

The first goal of the benchmarking activity is to identify examples of performances at each predefined performance level. One approach has been to use a developmental continuum similar to that developed by the National Board for Professional Teaching Standards (see for example Denner, Salzman & Bangert, 2001). Performances at the lowest level are categorized as Basic or *Beginning* level. Performances partially meeting the standards are be classified as *Developing*. Performances meeting all of the standards are classified as *Proficient*. Outstanding performances are placed in a separate category labeled *Exemplary*, so raters do not come to believe that only outstanding performances meet all of the standards (Denner, et al., 2001). Each performance category defines a level of performance in terms of an overall judgment of the degree to which the performance provides evidence of meeting all of the standards.

A benchmarking process can be conducted before or after the assessors have rated the collected performances. The process steps below illustrate the steps of a benchmarking process undertaken before detailed scoring using the assessment's rubric. One reason to conduct a benchmarking process before actual scoring takes place is to pre-sort the performances when a large number of performances are to be rated. Because performance assessments often take a long time to score, no one set of raters will be able to rate all of the performance in a limited time. As a result, the total set of performances must be divided into smaller sets. To insure each smaller set contains a representative range of performances, a stratified random sampling procedure can be used with the benchmark categorization serving as the stratifying factor. In this way, all groups of raters will rate comparable sets of performances. This makes estimation of inter-rater agreement better as will be discussed later in Component 9.

## **Process for this component:**

### **Step 1. Develop a Holistic Rubric**

Define holistic performance levels in terms of an overall judgment of the degree to which the performance provides evidence of meeting all of the standards. Develop a holistic rubric describing each performance level category. For example, as shown in Table 10, the Renaissance Partnership developed four levels of teaching performance: 1 - Beginning, 2 - Developing, 3 – Proficient, and 4 – Expert. The holistic rubric allows scorers to place performances into one of these four categories.

### **Step 2. Perform a Quick Read**

Divide your assembled group of raters into smaller groups of three or four raters. Ask each group to perform a *quick read* of a percentage (say 20%) of the collected performance assessments. The percentage will depend upon the number of rater groups, the number of performances, and the amount of time available for sorting the performances. Instruct the groups to reach consensus on the holistic score category and place each performance in one of the piles (on separate tables) representing the pre-designated holistic performance levels. Continuing the example above, scorers would place performances on table designated as 1, 2, 3, or 4 representing the Renaissance Partnership holistic categories.

### **Step 3. Choose Exemplars**

Later in the day, redistribute the raters into different small groups composed of three or four raters. Assign each group the task of picking a fixed number of exemplar performances from one of the holistic score category piles. Exemplars would be those that represent the *typical* performance in a holistic category—neither the *best* in the category nor the *worst*. In choosing exemplar performances, each group must reach consensus. Confirm the exemplar choices with detailed scoring using the assessment’s scoring rubric if time permits or obtain confirmatory ratings later using separate raters.

### **Expected Product of this Component:**

A set of “benchmarked” performance exhibits that are good examples of candidate performance at each of the levels defined by the rubric.

### **Tips and Good Advice for this Component:**

1. For several of the exemplars, marginalia or text-windowed comments can be generated to justify the ratings for use in the training of future raters.
2. During the benchmarking process, it is a good idea to have one or more assistants who can monitor the category piles to be sure the categorized performances are placed in the correct pile.
3. A laptop microcomputer can be used for on-the-spot data entry of the holistic categorizations.

**Table 10. Renaissance Teacher Work Sample Holistic Rubric**

Holistic Score \_\_\_\_\_

Rate the TWS overall using the following holistic scale:

<b>1 = Beginning</b>	<b>2 = Developing</b>	<b>3 = Proficient</b>	<b>4 = Expert</b>
----------------------	-----------------------	-----------------------	-------------------

**Beginning**

The Beginning performance provides *little or no evidence* of the teacher’s ability to plan, deliver, and assess a standards-based instructional sequence, analyze student learning, and reflect on his or her instruction and student learning to improve teaching practice.

**Developing**

The Developing performance provides *limited evidence* of the teacher’s ability to plan, deliver, and assess a standards-based instructional sequence, analyze student learning, and reflect on his or her instruction and student learning to improve teaching practice.

**Proficient**

The Proficient performance provides *sufficient evidence* of the teacher’s ability to plan, deliver, and assess a standards-based instructional sequence, analyze student learning, and reflect on his or her instruction and student learning to improve teaching practice.

**Expert**

The Expert performance provides *clear, consistent, and convincing evidence* of the teacher’s ability to plan, deliver, and assess a standards-based instructional sequence, analyze student learning, and reflect on his or her instruction and student learning to improve teaching practice.

## **Component 8: Scoring Performances**

**Key Question: What are the essentials for scoring performance assessments that produce credible results?**

**Task for this component:**

Conduct and manage a formal scoring session that produces credible performance ratings for all performance exhibits.

**Rationale for this component:**

Raters must be assigned the task of rating a set of performances. Multiple panels of qualified raters can be formed if a large number of performances are to be assessed. For credibility evidence studies, it is best if the raters are randomly assigned to a set of performances. Although raters can be allowed to take their assigned set of performance to other locations for evaluation, such as their offices, it is better if the raters are supervised during the rating process in a common location. In a supervised location, the evaluation process can proceed without the complication of other distractions that might cause scoring fluctuations. Data entry can also be speeded if the scores are entered into a database immediately upon return. Assembling all raters at a fixed time and location can also prevent evaluation delays associated with rater procrastination or interfering circumstances.

**Process for this component:**

**Step 1. Assign Raters and Groups**

After training, randomly assign the assembled raters to groups of four to six raters. Assigning each rater a number and then using a table or random numbers to assign the rater to one of a predetermined and fixed number of groups can accomplish this. The groups can be composed of mixed representatives such as cooperating teachers, program faculty members, and university supervisors. After groups are formed, also randomly assign each group to a separate set of performances.

**Step 2. Score Performances**

Each group of raters should be assigned to score as many performances as feasible, however fatigue can affect score accuracy and consistency. A minimum number of ten per rater should be used to obtain credibility evidence, but a larger number is better.

If a second day of scoring is required, take time to reorient the raters and remind them of the assessor guidelines. The actual scoring of performances should start after a complete review of the guidelines, the standards, and the directions for the scoring rubric. Also, remind raters to score independently and refer to their own *Hit List of Personal Biases*.

As the raters complete the scoring of each performance, direct them to submit their completed scoring rubric to a person assigned to enter the scores into a database. The raters can then return the finished performance back to the set and select another performance from the set to score. The raters continue this cycle until they have finished rating all the performances in their assigned set. In this way, the same candidate performances can be evaluated by multiple raters without having to make multiple copies. This procedure also helps to insure that the order in which performances are evaluated will not have a systematic influence on the scores assigned because the assigned raters will all rate the same performances but in a different order.

**Expected Product of this Component:**

A set of scores for all completed candidate exhibits that represents credible and reportable results of performance.

**Tips and Good Advice for this Component:**

1. It is a good idea to have someone available, such as the Assessment Coordinator, to advise raters about what to do with the “weird stuff”--performances that are missing required sections, contain irrelevant materials, or are difficult for the raters to read for various reasons. Someone must decide whether or not to proceed with the evaluation of the performance and should document the reasons for this decision.
2. As raters return completed rubrics, it is a good idea to have someone check them for completion. Raters can be asked to finish sections that were skipped before starting to rate another performance. This reduces the amount of missing data.

Draft

## Component 9: Determining Score Generalizability

**Key Question:** *What statistical processes are recommended to determine the inter-rater reliability in scoring performance assessments?*

**Task for this Component:**

Conduct formal studies of inter-rater agreement that provides statistical information about the dependability of judgments across different raters and scoring occasions.

**Rationale for this Component:**

It is important for scores on performance assessments to show a high degree of accuracy and consistency if the scores are going to be used for making high-stakes decisions about the performance levels of your teacher candidates. Hence, the judgments of the raters must be in close agreement with one another. It is also important to show the scores can be generalized beyond the particular tasks, the particular raters, and the particular occasion of assessment, if the scores are to be used to make general inferences about candidates' abilities to meet institutional and state teaching standards and their abilities to perform successfully as teachers. These issues can be examined for performance assessment ratings through the application of concepts from Generalizability Theory (see Shavelson & Webb, 1991).

Generalizability Theory (GT) provides a flexible alternative to Classical Test Theory by allowing multiple sources of score error to be estimated separately in the same research design. Generalizability Theory uses random effect analysis of variance (ANOVA) as a mechanism to distinguish between various sources of potential measurement error. Identified sources of measurement error such as tasks, occasions, or raters are called *facets*. Typically, facets are considered *random effects* because the levels of the facet (e.g., the different raters) either were chosen at random from a defined population or represent an interchangeable subset of raters from the population. The rater effect in the analysis of variance assesses the variability of scores on the performance assessment that are due to rater differences to determine whether the amount of variability is statistically significant. The *variance component* for the random effect of rater provides an estimate of the size of the variability in candidate scores that can be attributed to differences among raters from the population of raters that will be used to score their performances.

Additionally, Generalizability Theory permits the computation of a summary coefficient reflecting the level of dependability of the scores that is similar in interpretation to classical test theory's reliability coefficient (Shavelson & Webb, 1991). Once the required variance components have been calculated, they can be entered into formulas for computing coefficients of dependability. The key information for the necessary variance components is contained in the ANOVA printout. Dependability coefficients for scores used to make absolute decisions (criterion-referenced) about performance levels can be computed using formulas provided by Shavelson and Webb (1991). Single or multiple rater coefficients of dependability for absolute decisions can be computed by adjusting the number of raters included in the formula.

In addition to the rater facet, generalizability studies always include a *person facet*. The person facet is an estimate of the variation in scores due to differences in *candidate* performance levels. The variation reflects differences among the candidates in the knowledge and skills they

possess with respect to the competencies evaluated by the performance assessment. It is expected that candidates will vary in ability (Lunz & Schumacker, 1997, p. 220). In fact, higher coefficients of dependability are more likely to be obtained when there is a large spread in the scores of the candidates. As a result, all estimates of generalizability of scores are dependent upon the groups of candidates completing the performance assessment. Consequently, each institution must undertake its own credibility studies to establish score generalizability with respect to its own teacher education candidates and raters.

Because each performance assessment only provides an assessment of performance at a specific point in time, the generalizability of scores across performance occasions is also an important consideration. The *occasion facet* estimates variation with respect to the possible occasions on which a decision maker would be equally willing to accept a score on the performance assessment (Shavelson, Baxter, & Gao, 1993). The occasion facet also addresses the question of whether candidates would receive similar scores if asked to complete the same complex performance again. The *task facet* is concerned with the separate dimensions associated with overall performance. In other words, do the candidates perform consistently across the various tasks and performance dimensions rated on the scoring rubric? Are the candidates' performances generalizable across the separate tasks? Although the occasion facet and task facet are important considerations, the most important consideration is probably the generalizability of the candidates' scores across raters. Hence, research efforts to obtain credibility evidence are less likely to be concerned with either the occasion facet or the task facet until evidence has been obtained to support the generalizability of candidates' scores across raters. For an example of a credibility evidence study of teacher work sample assessment that included occasion as a facet see Denner, Salzman, Newsome, and Birdsong (2003).

### **Process for this Component:**

#### **Step 1. Enter Candidate scores into a Statistical Package and Compute an ANOVA**

To obtain the appropriate results, the easiest way to analyze the scores is to treat the effect of rater as a repeated-measures factor and conduct a single-factor repeated measures ANOVA. Table 11 presents the within-subjects test from an ANOVA printout generated using SPSS 10 for Macintosh (SPSS, 2000). Table 12 presents the between-subjects test from the same printout. Table 11 reveals the effect of rater on the total scores of teacher education candidates on the Renaissance Teacher Work Sample. For this analysis, six raters from different teacher education institutions scored the teacher work samples of ten teacher candidates collected from across different teacher education programs participating in the Renaissance partnership project (Denner, Norman, Salzman, Pankratz and Evans, in press). A single-factor repeated measures ANOVA was performed, so there is no between-subjects effect in this analysis. However, the mean square error shown in Table 12 represents the variance between persons and this information is needed to calculate a variance component used in the equation for determining a dependability coefficient. As can be seen from Table 11, the effect for rater was not statistically significant in this instance. This is good news. However, because complex performance assessments require the application of professional judgment during rating, a finding of no significant scoring differences among the raters is not to be expected. Rather it is the extent of the differences and the dependability of the score decisions made by the raters that is the important consideration. The computation of coefficients of dependability will be addressed in the next step.

**Table 11. Example SPSS ANOVA Printout for Test of Within-Subjects Effects**

Tests of Within-Subjects Effects						
Measure: MEASURE_1						
Source		Type III Sum of Squares	df	Mean Square	F	Sig.
RATERS	Sphericity Assumed	539.083	5	107.817	1.068	.391
	Greenhouse-Geisser	539.083	1.745	308.948	1.068	.358
	Huynh-Feldt	539.083	2.129	253.162	1.068	.367
	Lower-bound	539.083	1.000	539.083	1.068	.328
Error(RATERS)	Sphericity Assumed	4542.083	45	100.935		
	Greenhouse-Geisser	4542.083	15.704	289.229		
	Huynh-Feldt	4542.083	19.165	237.004		
	Lower-bound	4542.083	9.000	504.676		

**Table 12. Example SPSS ANOVA Printout for Test of Between-Subjects Effects**

Tests of Between-Subjects Effects					
Measure: MEASURE_1					
Transformed Variable: Average					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	323106.817	1	323106.817	419.195	.000
Error	6937.017	9	770.780		

**Step 2. Compute Estimates of Variance Components**

For the first step in computing generalizability coefficients, compute estimates of the variance components used in the formulas. For a single facet study examining the effect of rater, you will need to compute three variance components—person, rater, and residual error. The first variance component estimates the variability of the scores of the persons who completed the performance assessment. To compute this variance component, the residual error mean square from Table 11 is subtracted from the mean square between subjects from Table 12 and divided by the number of raters (in this case, 6 raters) included in the study as shown in the following formula.

$$\sigma_p^2 = \frac{MS_p - MS_{Res}}{n_R} = \frac{770.780 - 100.935}{6} = 111.641$$

In like manner, the variance component for the raters is calculated by subtracting the residual error mean square (shown in Table 11) from the mean square for raters (also shown in Table 11) and divided by the number of performance scores (in this case, 10 TWS). This equation is shown below.

$$\sigma_R^2 = \frac{MS_R - MS_{Res}}{n_p} = \frac{107.817 - 100.935}{10} = .688$$

The third variance component is the residual error mean square itself. It is represented symbolically by the following expression.

$$\sigma_{\text{Res}}^2 = 100.935$$

### Step 3. Compute Dependability Coefficients

To compute the dependability coefficient, divide the variance component for persons (candidates) by itself plus the variance component for making absolute (criterion-referenced) decisions. In this case, the variance component for absolute decisions is the variance component for raters plus the variance component for residual error divided by the number of raters who evaluated this set of performances. The formula is shown below along with an example.

$$\phi_6 = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_{\text{Abs}}^2} = \frac{\sigma_P^2}{\sigma_P^2 + (\sigma_R^2 + \sigma_{\text{Res}}^2)/n_R}$$

$$\phi_6 = \frac{111.641}{111.641 + (.688 + 100.935)/6} = \frac{111.641}{111.641 + 16.937} = \frac{111.641}{128.578} = .868$$

Like reliability coefficients, dependability coefficients are usually reported to two decimal places. Thus, in this case the six-rater dependability for scoring Renaissance Teacher Work Samples would be estimated to be .87. This is a very good dependability coefficient, indicating a high proportion of the score differences among teacher candidates is generalizable across raters.

### Step 4. Adjust Formulas for Number of Ratets

Adjust the formula for computing a dependability coefficient to determine the degree of dependability that would be achieved if only a single rater were used to assess the performances (Shavelson & Webb, 1991). The formula is shown below along with a completed example.

$$\phi_1 = \frac{\sigma_P^2}{\sigma_P^2 + (\sigma_R^2 + \sigma_{\text{Res}}^2)}$$

$$\phi_1 = \frac{111.641}{111.641 + (.688 + 100.935)} = \frac{111.641}{111.641 + 101.623} = \frac{111.641}{213.264} = .523$$

To determine the minimum number of raters necessary for making high-stakes decisions about candidates' performance levels, adjust the formula to estimate dependability coefficients for panels of raters composed of different numbers of raters. This is done by adjusting the number of raters included in the formula shown in Step 3 above. The illustration below is for three raters based on the data from the Denner et al. (2004) investigation.

$$\phi_3 = \frac{111.641}{111.641 + (.688 + 100.935)/3} = \frac{111.641}{111.641 + 33.874} = \frac{111.641}{145.515} = .767$$

Again, the dependability coefficient of .77 indicates a high proportion of the score differences among teacher candidates are generalizable from a panel of only three raters. If .75 is taken as the criterion for an acceptable level of dependability for making decisions about candidate performances, then the data from this investigation support the generalizability of scores on the Renaissance Teacher Work Sample when the performance are rated by panels of three or more judges.

**Expected Product of this Component:**

A dependability coefficient for each group of raters that scored a set of performance exhibits.

**Tips and Good Advice for this Component:**

1. Keep in mind when planning a generalizability study that although it is difficult for the same panel of judges to score a large number of performance assessments, having each panel score only a small number of performances is likely to decrease the magnitude of a dependability coefficient (the same is true for all reliability coefficients). Thus, it is desirable to have the judges score as many performances as feasible. The example shown had only ten teacher work sample performances. This number is quite small for maximizing the magnitude of a dependability coefficient. It should be considered a minimum number rather than an optimal number of performances.
2. For the analysis of complex generalizability studies that include additional facets such as tasks or occasions see the primer by Shavelson and Webb (1991) or consult a statistician.

## Component 10: Gathering Validity Evidence

**Key Question:** *What types of evidence should be collected to establish the validity of performance assessments?*

**Tasks for this Component:**

Gather, analyze and report judgments of credible professionals about the alignment of your performance assessments with teaching standards and their validity using criteria recommended by Crocker (1997).

**Rationale for this Component:**

Validity evidence is essential for any performance assessment to be credible. Validity asks the question, does the assessment measure what it purports to measure? Performance assessments should not be used to make decisions about candidate performance levels until after adequate credibility evidence has been collected supporting use of the assessment for that purpose (Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, 1999). Perhaps the most important aspect of validity for a performance assessment is whether successful performance depends on the knowledge and skills targeted by the assessment. In other words, do the tasks elicit performances that are representative of the standards? For teacher education candidates, the knowledge and skills to be assessed are those reflected in the institutional and state standards that are the target of the assessment. The tasks the candidates are asked to perform and the indicators used to judge the task performances must directly measure the knowledge and skills entailed in the standards. Conversely, although a single, complex performance assessment may address multiple standards, it is unlikely to measure all of your institutional and state teaching standards. Therefore, it is also valuable to show that the tasks required by the performance assessment are not judged to assess non-targeted standards. The match to the targeted standards should be high and direct, but the tasks should not be judged to match standards the assessment was not designed to assess. At the same time, it is important to show the tasks do not require extraneous knowledge and skills that are unrelated to the targeted standards. The latter is also an issue of fairness (discussed separately elsewhere). Ideally, the corresponding elements of the performance tasks, the scoring rubric and the targeted standards will be judged to have high alignment because any discrepancies are cause for concern about the validity of the assessment (Crocker, 1997).

In addition to showing alignment with the targeted standards, it is also desirable to demonstrate that the task performances required by the assessment are representative of the broader assessment domain (i.e., actual teaching performance). Crocker (1997) has proposed criteria for judging the content representativeness of performance assessments that include dimensions such as the *criticality* of the tasks to actual teaching performance, the *frequency* with which the tasks would be performed during actual teaching, and the *realism* of the performance tasks as simulations of actual classroom performances. Another important criterion is *balance* because the candidates only complete a small number of performance tasks. It is, therefore, desirable for the tasks to reflect well the entire performance domain. Typically, a rating scale is used to

assess a criterion such as *criticality* or *realism* along a continuum from 1 = not at all important (or realistic) to 4 = very important (or very realistic). The criterion of *frequency* is usually assessed by asking the panel of raters to determine how often the teacher candidate would be expected to perform the required tasks during the course of actual teaching. Thus, summary ratings from a representative panel of raters who independently affirm these dimensions expresses the consensus of a professional community that the results of the assessment reveal something significant about teacher candidates' abilities to teach. It is important to remember, however, that validity is an ongoing argument that combines both logical and empirical elements, and that it does not refer to the assessment itself but only to the use of its results for certain purposes. Hence, the examples presented here are not exhaustive of the array of validity evidence that could be gathered in support of the use of a particular performance assessment for its intended purposes.

**Process for this Component:**

**Step 1. Develop Performance Alignment Scales**

Develop a rating scale that asks the validity assessment panel to judge the overall alignment among the tasks specified in the assessment guidelines, the indicators specified in the scoring rubric, and the standards targeted by the performance assessment. Table 13 provides an example scale similar to the one used to assess the overall alignment of the Renaissance Teacher Work Sample guidelines, scoring rubric, and targeted teaching process standards (Denner et al., 2004). Table 13 presents an example of a scale that was used to rate the degree to which the tasks specified in the Renaissance Teacher Work Sample were considered to be representative of the targeted teaching process standards, and thus to be well aligned with those standards.

**Table 13. Example Overall Alignment Scale**

<i>Directions:</i> Indicate the overall degree to which the elements of the Renaissance Teacher Work Sample guidelines, and the scoring rubric are <i>aligned</i> with the targeted teaching process standards and with each other by checking or marking with an X the appropriate space in the table below.				
Alignment Comparisons	Degree of Alignment			
	Poor 1	Low 2	Moderate 3	High 4
Alignment of the Renaissance Teacher Work Sample guidelines and task prompts with the targeted teaching process standards.				
Alignment of the Renaissance Teacher Work Sample guidelines and task prompts with the scoring rubric.				
Alignment of the scoring rubric with the targeted teaching process standards and indicators.				

**Table 14. Example Task Representativeness Scale**

*Directions:* Indicate the degree to which the tasks required by the Teacher Work Sample *align* with and are *representative* of the targeted teaching process standards by marking the appropriate space below.

Tasks Required By the Teacher Work Sample	Degree of Representativeness			
	Not at all Representative 1	Somewhat Representative 2	Representative 3	Very Representative 4
Teacher uses understanding of student individual differences and community, school, and classroom characteristics to draw specific implications for instruction and assessment.				
Teacher sets significant, challenging, varied, and appropriate learning goals for student achievement that are aligned with local, state, or national standards.				
Teacher designs an assessment plan to monitor student progress toward learning goals, using multiple assessment modes and approaches to assess student learning before, during, and after instruction.				
Teacher designs instruction aligned to learning goals and with reference to contextual factors and pre-assessment data, specifying instructional topics, learning activities, assignments and resources.				
Teacher designs instruction with content that it accurate, logically organized, and congruent with the big ideas or structure of the discipline.				
Teacher uses on-going analysis of student learning and responses to rethink and modify original instructional design and lesson plans to improve student progress toward the learning goal(s).				
Teacher analyzes assessment data, including pre/post assessments and formative assessments, to determine students' progress related to the unit learning goals.				
Teacher uses graphs or charts to profile whole class performance on pre-assessments and post-assessments, and to analyze trends or differences in student learning for selected subgroups.				
Teacher evaluates the effectiveness of instruction and reflect upon teaching practices and their effects on student learning, identifying future actions for improved practice and professional growth				

## **Step 2. Develop Performance/Standard Alignment Rating Scales**

Develop additional rating scales that ask the panel members to match the task elements of the performance and the scoring rubric to institutional and state performance standards. Table 15 provides an example rating scale similar to the one used by Denner, et al. (2004) to assess the alignment of the tasks required by the Renaissance Teacher Work Sample to the INTASC standards (Interstate New Teacher Assessment and Support Consortium, 1992). Because many states have developed beginning teacher licensure standards based on the INTASC standards and many institutional standards also reflect the INTASC standards, this scale can be easily adapted by your institution for use with any of your performance assessments of teacher candidates. In the example, the rating scale of 1 = *Not at All*; 2 = *Implicitly*; and 3 = *Directly* was used to determine whether the tasks performances required by the Renaissance Teacher Work Sample were seen to be a direct measure of any of the INTASC standards. Other rating dimensions could have been employed instead, such as the degree of alignment scale shown in Table 13 or the degree of representativeness scale shown in Table 14. Likewise, following a similar pattern, additional scales could be developed to compare the degree of alignment of the specific task elements of your performance assessment to their related scoring rubric indicators. The number of rating scales you choose to employ will depend upon your performance assessment and the level of detailed analysis you desire with respect to alignment. Credibility evidence for alignment can vary from global confirmation to detailed verification.

**Table 15. Example Alignment with Standards Scale**

<i>Directions:</i> Indicate the extent to which the task performances of the Renaissance Teacher Work Sample measure the Interstate New Teacher Assessment and Support Consortium (INTASC) model standards for beginning teachers by marking the appropriate space in the table below.			
INTASC Standards	Degree of Alignment		
	Not at all 1	Implicitly 2	Directly 3
<p><b><u>Knowledge of Subject Matter</u></b>  <i>The teacher understands the central concepts; tools of inquiry, and structures of the content area(s) taught and create learning experiences that make these aspects of subject matter meaningful for learners.</i></p>			
<p><b><u>Knowledge of Human Development and Learning</u></b>  <i>The teacher understands how students learn and develop, and provides opportunities that support their intellectual, social, and personal development.</i></p>			
<p><b><u>Adapting Instruction for Individual Needs</u></b>  <i>The teacher understands how students differ in their approaches to learning and creates instructional opportunities that area adapted to learners with diverse needs.</i></p>			
<p><b><u>Multiple Instructional Strategies</u></b>  <i>The teacher understands and uses a variety of instructional strategies to develop students' critical thinking, problem solving, and performance skills.</i></p>			
<p><b><u>Classroom Motivation and Management Skills</u></b>  <i>The teacher understands individual and group motivation and behavior and creates a learning environment that encourages positive social interaction, active engagement in learning, and self-motivation.</i></p>			
<p><b><u>Communication Skills</u></b>  <i>The teacher uses a variety of communication techniques including verbal, nonverbal, and media to foster inquiry, collaboration, and supportive interaction in and beyond the classroom.</i></p>			
<p><b><u>Instructional Planning Skills</u></b>  <i>The teacher plans and prepares instruction based upon knowledge of subject matter, students, the community, and curriculum goals.</i></p>			
<p><b><u>Assessment of Student Learning</u></b>  <i>The teacher understands, uses, and interprets formal and informal assessment strategies to evaluate and advance student performance and to determine program effectiveness.</i></p>			
<p><b><u>Professional Commitment and Responsibility</u></b>  <i>The teacher is a reflective practitioner who demonstrates a commitment to professional standards and is continuously engaged in purposeful mastery of the art and science of teaching.</i></p>			
<p><b><u>Partnerships</u></b>  <i>The teacher interacts in a professional, effective manner with colleagues, parents, and other members of the community to support students' learning and well-being.</i></p>			

### Step 3. Develop Validity Rating Scales

Establish a set of validity criteria for the performance assessment and develop a rating scale for each criterion. Table 16 shows an example rating scale using Crocker’s (1997) criterion of *criticality* (or importance). The example scale is the same as one employed by Denner et al. (2004) to assess the importance of the teaching behaviors measured by the Renaissance Teacher Work Sample to success as a classroom teacher. Table 17 presents an example for the criterion of *frequency* and Table 18 presents an example for the criterion of *authenticity* (or realism).

You will also want to develop questions and rating criteria that address the overall validity and use of your performance assessment as a measure of teacher candidate proficiency and as a measure of your institutional and state teaching standards. Table 19 presents an example set of questions and the rating criteria that were used by Denner et al. (2004) to assess the validity of the Renaissance Teacher Work Sample as a whole. The questions are illustrative of the kinds of questions that could be asked about any teacher candidate performance assessment.

**Table 16. Example Importance Rating Scale**

<i>Directions:</i> Please rate the <i>importance</i> of the teaching behaviors measured by the Renaissance TWS to success as a classroom teacher by checking or marking with an X the appropriate space in the table below.				
Teaching Behaviors Targeted By Teacher Work Sample	Degree of Importance			
	Not at all Important 1	Somewhat Important 2	Important 3	Very Important 4
Use information about the learning-teaching context and student individual differences to set learning goals and plan instruction and assessments.				
Set significant, challenging, varied, and appropriate learning goals.				
Use multiple assessment modes and approaches aligned with learning goals to assess student learning before, during, and after instruction.				
Design instruction for specific learning goals, student characteristics and needs, and learning contexts.				
Use ongoing analysis of student learning to make instructional decisions.				
Use assessment data to profile student learning and communicate information about student progress and achievement.				
Reflect on instruction and student learning in order to improve teaching practice.				

**Table 17. Example Frequency Rating Scale**

*Directions:* Please indicate *how frequently* you would expect a beginning teacher to engage in each of the following teaching behaviors measured by the Renaissance TWS during the course of his or her professional practice by checking or marking with an X the appropriate space in the table below.

Teaching Behaviors Targeted By Teacher Work Sample	Frequency				
	Never	Yearly	Monthly	Weekly	Daily
Use information about the learning-teaching context and student individual differences to set learning goals and plan instruction and assessments.					
Set significant, challenging, varied, and appropriate learning goals.					
Use multiple assessment modes and approaches aligned with learning goals to assess student learning before, during, and after instruction.					
Design instruction for specific learning goals, student characteristics and needs, and learning contexts.					
Use ongoing analysis of student learning to make instructional decisions.					
Use assessment data to profile student learning and communicate information about student progress and achievement.					
Reflect on instruction and student learning in order to improve teaching practice.					

**Table 18. Example Authenticity Rating Scale**

<i>Directions:</i> Please indicate how <i>authentic</i> the tasks required by the Renaissance TWS are to success as a classroom teacher by checking or marking with an X the appropriate space below.				
Tasks Required By the Teacher Work Sample	Degree of Authenticity			
	Not at all Authentic 1	Somewhat Authentic 2	Authentic 3	Very Authentic 4
Teacher uses understanding of student individual differences and community, school, and classroom characteristics to draw specific implications for instruction and assessment.				
Teacher sets significant, challenging, varied, and appropriate learning goals for student achievement that are aligned with local, state, or national standards.				
Teacher designs an assessment plan to monitor student progress toward learning goals, using multiple assessment modes and approaches to assess student learning before, during, and after instruction.				
Teacher designs instruction aligned to learning goals and with reference to contextual factors and pre-assessment data, specifying instructional topics, learning activities, assignments and resources.				
Teacher designs instruction with content that is accurate, logically organized, and congruent with the big ideas or structure of the discipline.				
Teacher uses on-going analysis of student learning and responses to rethink and modify original instructional design and lesson plans to improve student progress toward the learning goal(s).				
Teacher analyzes assessment data, including pre/post assessments and formative assessments, to determine students' progress related to the unit learning goals.				
Teacher uses graphs or charts to profile whole class performance on pre-assessments and post-assessments, and to analyze trends or differences in student learning for selected subgroups.				
Teacher evaluates the effectiveness of instruction and reflect upon teaching practices and their effects on student learning, identifying future actions for improved practice and professional growth				

**Table 19. Example Overall Validity Questions**

<i>Directions:</i> Please answer the following items considering the Renaissance Teacher Work Sample as a <i>whole</i> , by circling the one response that most closely reflects your judgment.			
A. Overall, does the Renaissance Teacher Work Sample measure knowledge and skills that are <i>necessary</i> for a beginning teacher?			
Not at All Necessary	Somewhat Necessary	Necessary	Absolutely Necessary
1	2	3	4
B. Overall, <i>how critical</i> to the practice of a beginning teacher are the teaching competencies the Renaissance Teacher Work Sample requires teacher candidates to demonstrate?			
Not at all Critical	Somewhat Critical	Critical	Absolutely Critical
1	2	3	4
C. Overall, does the Renaissance Teacher Work Sample present teacher candidates with <i>realistic</i> performance tasks similar to ones they might enact in professional practice as a teacher?			
Not at all Realistic	Somewhat Realistic	Realistic	Absolutely Realistic
1	2	3	4
D. Overall, how <i>appropriate</i> is it to use the Renaissance Teacher Work Sample as one measure of a teacher candidate's performance level with respect to beginning teacher standards?			
Not at all Appropriate	Somewhat Appropriate	Appropriate	Absolutely Appropriate
1	2	3	4

**Step 4. Develop the Validity Questionnaire**

After creating rating scales similar to those described above, organize them into a validity questionnaire.

**Step 5. Choose and Prepare the Validity Assessment Panel**

Your validity assessment panel may include a mix of teacher education faculty members, arts and science faculty members, and practicing educators. Other constituencies could also be represented on the panel, such as business and community leaders or current teacher education candidates. Separate panels could also be considered, such as a panel composed only of cooperating teachers from partnership schools. The size of the panel is less important than the degree to which the members are representative of the various constituencies. Recruitment of larger panels in the range of 40 to 50 raters may afford an opportunity to demonstrate wide spread agreement and support for the assessment.

Acquaint the validity assessment panel with the performance assessment guidelines, scoring rubric, and the targeted standards. The panel members must have a complete acquaintance with the performance tasks and the performance indicators used to judge the candidates' performances before they can be expected to judge their alignment to each other or to the targeted standards. At the very least, the panel members should be given an opportunity to examine examples of actual teacher candidate performances on the assessment before they are asked to judge the various aspects of alignment. Ideally, the panel members would also have had some experience using the scoring rubric to score candidate performances.

### **Step 6. Collect and Analyze Validity Results**

Collect the validity questionnaires and summarize the ratings. Use a frequency count of the number of validity panel members who chose each rating for each validity consideration. The frequencies may also be converted to percentages. Ideally, the frequencies and percentages will show support for high alignment, for representative assessment of the targeted standards, and for the other validity criteria addressed in your validity questionnaire.

#### **Expected Product of this Component:**

A set of data including statistical results of validity judgments by credible professionals for each validity factor considered and tested.

#### **Tips and Good Advice for this Component:**

1. The panel members can be allowed a means to express their opinions through open-ended questions or comments (Crocker, 1997). Leave space below the rating scales for comments by inserting the word "Comments:" and modifying the instructions to inform the panel members that their comments are welcome. Comments are very useful during the development stages of a performance assessment when some of the alignments may be weak. The comments can provide useful insights about needed improvements.
2. Smaller panels of raters can be used during the development stages of your performance assessment with larger panels reserved to gather validity evidence after the development work is complete. It usually takes several iterations before tight alignment is achieved among the performance assessment tasks specified in the guidelines, the indicators on the scoring rubric, and the targeted performance standards.
3. In addition to summarizing the responses for each of Crocker's (1997) criteria, it is also of benefit to inspect the "pattern of correlation among the criteria" (p. 90).

## **Component 11: Developing Your Data Collection Timeline & Initial Reports**

**Key Question:** *What steps should be taken to ensure that collecting, analyzing, summarizing, and reporting of assessment results become part of a routine “culture of evidence”?*

**Task for this Component:**

Identify key questions that the data system must answer and the audiences who will benefit from the results of data analysis. Develop a timeline to gather, analyze, and summarize assessment data as they relate to targeted teaching standards and to report these results in a form that meets the needs of various audiences. Develop initial reports.

**Rationale for this Component:**

As we live in an information and data management age, all institutions expend much time and energy collecting multiple pieces of data about their students into one or more data systems. A tragic flaw in most systems, however, is that institutions have never considered the purpose for the data they collect except as a response to external demands from institutional hierarchies and state and federal organizations to provide sometimes seemingly arbitrary information. Although these external organizations cannot be ignored, it is arguable that data collection becomes an end in itself versus serving the role of informing institutions about their progress toward self-selected goals. To maximize the benefits of data collection and management, institutions must identify goals and outcomes, generate key questions related to the outcomes that collected (or potentially collectible) data can answer, identify key audiences and constituents to report findings, and set a timeline to ensure that data are collected, analyzed, and reported in such a manner to allow for making essential decisions related to institutional goals. Thus, in each institution, an accountability culture must prevail that guides the data management system.

Furthermore, it is important to keep in mind that assessment data must be reported as they relate to teaching standards and/or goals articulated in your conceptual framework and other mission documents. Reports that are divorced from these standards and goals are likely to be ignored as irrelevant by their audience. Reports that meet with audience indifference should be re-evaluated as to whether they are disconnected from targeted standards or goals or whether the intended audience needs further education regarding the importance of the data in the report.

Finally, if the relationship between performance assessments and targeted teaching standards are not readily apparent, then we suggest that you revisit earlier components of this Credibility Manual. It is also important that institutions have done the necessary work associated with earlier components of this Manual to establish that assessments are reliable measures of candidate performance and are considered valid measures of teacher standards.

## **Process for this Component:**

### **Step 1. Identify Institutional Outcomes and Related Key Questions**

Identify student and program outcomes that will represent success in your institutions and programs. Besides the work begun in Component 1 of this manual, you will also want to consult your existing institutional conceptual framework and/or mission statement. Also, most institutions will rely on national and/or state standards (or SPAs) as a beginning list for outcomes, but many may expand upon these based on particular institutional values that they hope to instill. By way of example, the foundation for WKU's conceptual framework is student progress and success in meeting Kentucky's Teacher Standards.

After identifying outcomes, generate key questions that the performance assessments and accountability system must answer. Below is a list of sample questions, based on WKU's goal of meeting Kentucky's Teacher Standards.

1. What is the performance level of teacher candidates at any point during preparation and exit relative to one or more teacher standards?
2. What is the progress of students relative to a performance assessment (tied to teacher standards) as a result of completing a course or set of learning experiences?
3. Which teacher standards are addressed well/poorly by the performance assessment system?
4. What is the distribution of performance levels of all candidates at exit from the preparation program?
5. What is the contribution of particular candidate demographic and entrance factors to teacher performance at exit and performance on the job?
6. What are candidate and/or graduates' perceptions regarding their preparation to meet each of the state teacher standards?
7. What performance scores related to Kentucky's Teacher Standards are the best predictors of a new teacher's ability to facilitate learning for all students they teach?

### **Step 2: Identify Audiences for Assessment Results Reports**

Identify, in collaboration with colleagues from the various departments or schools that make up the teacher education unit, the audiences who will receive and make decisions based on the collected assessment data. A good place to start would be to review documents of your accrediting agency to see what it suggests as intended audiences for reporting data. For example, a look at the NCATE 2002 standards reveals that this agency expects assessment data to be reported to students, faculty, programs, administration, learned societies, and to NCATE itself. Table 20 shows some of WKU's primary audiences and what they need to know.

**Table 20: Table of Primary Audiences (WKU Example)**

Report Audiences	What They Need to Know
Teacher Candidates	<ul style="list-style-type: none"> <li>• Individual progress toward teacher standards</li> <li>• Areas of strengths and weaknesses related to teacher standards</li> </ul>
University Faculty	<ul style="list-style-type: none"> <li>• Performance of students in a class relative to other classes</li> <li>• Progress reports on students they advise</li> </ul>
Program Administrators & Curriculum Committees	<ul style="list-style-type: none"> <li>• Overall program performance relative to moving candidates toward standards</li> <li>• Strengths and weaknesses of program elements</li> <li>• Candidate perceptions of preparation relative to standards</li> </ul>
NCATE (or Other Accrediting Agencies)	<ul style="list-style-type: none"> <li>• Progress toward self-identified and NCATE established standards</li> </ul>
School Practitioners & Employers	<ul style="list-style-type: none"> <li>• Quality of graduates relative to standards</li> </ul>
Policy Makers	<ul style="list-style-type: none"> <li>• Performance of graduates relative to standards</li> <li>• Productivity of programs</li> <li>• Cost effectiveness of preparation programs</li> </ul>

As can be seen, not all constituents need the same data. For example, a candidate most likely will look to assessment data to answer the question, “How am I progressing toward meeting the requirements for graduation?” Thus, reports to candidates will need to focus on their ability to meet teaching standards with the goal of providing guidance for continued progress. Similarly, reports to faculty should focus on how assessment data help guide decisions about changes to improve programs. Reports to administration should reflect adequacy/inadequacy of resources/personnel to move students towards proficiency on teacher standards.

**Step 3. Identify Data to Enter in the Institutional Database**

One of the greatest challenges is identifying what data are needed in order to answer the key questions. Of course, scores on performance assessments should be entered. However, it may take several iterations to create a fully functioning data management system that contains adequate data to all key questions.

It should be emphasized that some care should be exercised in entering data into the institutional database to ensure their easy analyses. For example, the data management system should be set up to easily aggregate performance assessments of a particular student as they relate to progress toward particular teaching standards and/or program goals. Likewise, the database should permit the summarization of all candidate performances by program and/or by teacher standard. Ensuring that these data can “talk to each other” would involve discussion with a computer programmer (beyond the scope of this manual).

#### **Step 4. Develop an Assessment Data Collection and Reporting Timeline**

Develop a timeline that identifies the assessments to be carried out and the frequency of the assessment. Additional columns in the timeline may include initiation date, report date, person responsible, and primary report audience. Identify the assessments down the side and the frequency and other important items across the top. Table 21 provides an example timeline from Western Kentucky University.

With the help of colleagues in both faculty and administration, review each of the assessments and attempt to determine such information as how often they will be administered, by whom, the anticipated date by which data will be analyzed and reported, and who the intended audiences may be for the reports. In setting deadlines in the timeline for reporting data, it is important to remember that they must reflect sufficient time for program and administration to make timely decisions based on the report.

#### **Step 5. Report Results based on Key Questions**

On the following pages (Tables 22-24) are sample reports generated to answer key questions identified earlier in Table 21. Table 22 illustrates how students scored on WKU's Critical Performances related to New Teacher Standard 1 during one semester. For each course in the WKU teacher education program, faculty members who teach that course have worked together to create assessments called critical performance that address one or more New Teacher Standards. Any course may have one or more critical performances (CPs). For example, in Table 22, the course CFS 191 has two CPs. Using a rubric format similar to the Teacher Work Sample, students are graded as 1 – Beginning, 2 – Developing, 3 – At Standard, or 4 – Above Standard. Furthermore, WKU CPs are leveled in terms of the depth of knowledge they were created to assess. There are four knowledge levels: Level I - Knowledge/Awareness, Level II - Developing/ Beginning Application, Level III - Application/ Analysis, and Level IV - Synthesis/Evaluation. Thus, any CPs listed above the "1 Total" row on Table 22 are Level 1 – Knowledge/Awareness performances and, on these performances students can score (CPSCORE on Table 22) from 1 – Beginning to 4 – Above Standard; those CPs above the "2 Total" row are Level II – Developing/Beginning Application, and so on. Because faculty creating these CPs explicitly identify which New Teacher Standards the CPs measure, it becomes relatively easy to use Excel to identify how students are scoring on only those CPs related to each standard. Although not shown here, within the Excel file from which Table 22 was created are pages like this one for each New Teacher Standard. A final feature of this table is that the first set of columns tells the number of students performing at each level whereas the second set of columns translates the numbers into percentages. Thus, reading Table 22 reveals, for example, the following about students completing CP 2 in CFS 191: 11 (17%) scored 1 – Beginning, 8 (13%) scored 2 – Developing, 28 (44%) scored 3 – At Standard, and 17 (27%) scored 4 – Above Standard.

**Table 21. Assessment Timeline (WKU Example)**

Key Questions	Assessments and Studies	Schedule	Initiation Date	Report Date	Person Responsible
1. Are teacher candidates making progress towards meeting Kentucky Teacher Standards?	Critical Performances	Every Semester	Note: New CPs must be entered into system 1 month prior to new semester.	March (Fall) June (Spring)	Data System Manager
2. Has the preparation program adequately prepared students to demonstrate proficiency in the performance Teacher Standards?	Teacher Work Sample	Every Semester		March (Fall) June (Spring)	Data System Manager
3. How prepared to meet Kentucky Teacher Standards do student-teachers feel?	Student-Teacher Survey	Every Semester	Nov. - Fall April – Spring	June	University Supervisors
4. How prepared to meet Kentucky Teacher Standards do first year teachers feel?	First Year Teacher Survey	Yearly	February 1 – First Wave March 1 – Second Wave	April	Assessment Coordinator
5. How prepared to meet Kentucky Teacher Standards do second year teachers feel?	Second Year Teacher Survey	Yearly	February 1 – First Wave March 1 – Second Wave	April	Assessment Coordinator
6. To what extent do employers feel recent graduates are prepared to meet Kentucky Teacher Standard?	Principal Survey	Yearly	March 1 – First Wave April 1 – Second Wave	May	Assessment Coordinator
7. Can faculty consistently differentiate among various levels of candidate performance?	TWS Inter-rater Reliability	Yearly	Summer	August	Faculty
		<b>SPECIAL CONSIDERATIONS:</b> Yearly calibration meeting for returning faculty Yearly training meeting for new faculty			
8. Can faculty consistently differentiate among various levels of candidate performance?	Critical Performance Inter-rater Reliability	On-going	Fall, 2004	Ongoing	Faculty
9. What program changes and decisions have been made based on assessment data?	Program Changes based on TWS, CPs, and Surveys	<b>SPECIAL CONSIDERATIONS:</b> Need to develop a SYSTEMATIC reporting form to document these changes as they occur.			Department Heads

**Table 22. Critical Performance Data to answer Key Question 1 (from Table 20) – New Teacher Standard 1 (WKU Example)**

**Critical Performance By New Teacher Standard 1 by Level (Count and Percent)**

Score Count			CPSCORE					CPSCORE				
NTS 1	COURSE	CPNO	1	2	3	4	Grand Total	1	2	3	4	Grand Total
1	CFS-191	1	1	6	10	5	22	5%	27%	45%	23%	100%
		2	11	8	28	17	64	17%	13%	44%	27%	100%
	<b>CFS-191 Total</b>		<b>12</b>	<b>14</b>	<b>38</b>	<b>22</b>	<b>86</b>	<b>14%</b>	<b>16%</b>	<b>44%</b>	<b>26%</b>	<b>100%</b>
	CFS-192	3		1	1	21	23	0%	4%	4%	91%	100%
	<b>CFS-192 Total</b>			<b>1</b>	<b>1</b>	<b>21</b>	<b>23</b>	<b>0%</b>	<b>4%</b>	<b>4%</b>	<b>91%</b>	<b>100%</b>
	EDU-250	1	3	7	113	114	237	1%	3%	48%	48%	100%
		2	3	9	114	102	228	1%	4%	50%	45%	100%
	<b>EDU-250 Total</b>		<b>6</b>	<b>16</b>	<b>227</b>	<b>216</b>	<b>465</b>	<b>1%</b>	<b>3%</b>	<b>49%</b>	<b>46%</b>	<b>100%</b>
	ELED-355	1		1	96	23	120	0%	1%	80%	19%	100%
		2			212	20	232	0%	0%	91%	9%	100%
		3			462	12	474	0%	0%	97%	3%	100%
	<b>ELED-355 Total</b>			<b>1</b>	<b>770</b>	<b>55</b>	<b>826</b>	<b>0%</b>	<b>0%</b>	<b>93%</b>	<b>7%</b>	<b>100%</b>
	LME-288	1			25	15	40	0%	0%	63%	38%	100%
	<b>LME-288 Total</b>				<b>25</b>	<b>15</b>	<b>40</b>	<b>0%</b>	<b>0%</b>	<b>63%</b>	<b>38%</b>	<b>100%</b>
MGE-275	1	1		1	72	74	1%	0%	1%	97%	100%	
	2	1			70	71	1%	0%	0%	99%	100%	
	3				63	63	0%	0%	0%	100%	100%	
	4			2	64	66	0%	0%	3%	97%	100%	
<b>MGE-275 Total</b>		<b>2</b>		<b>3</b>	<b>269</b>	<b>274</b>	<b>1%</b>	<b>0%</b>	<b>1%</b>	<b>98%</b>	<b>100%</b>	
<b>1 Total</b>			<b>20</b>	<b>32</b>	<b>1064</b>	<b>598</b>	<b>1714</b>	<b>1%</b>	<b>2%</b>	<b>62%</b>	<b>35%</b>	<b>100%</b>
2	ELED-345	1			158		158	0%	0%	100%	0%	100%
		3		5	222		227	0%	2%	98%	0%	100%
	<b>ELED-345 Total</b>			<b>5</b>	<b>380</b>		<b>385</b>	<b>0%</b>	<b>1%</b>	<b>99%</b>	<b>0%</b>	<b>100%</b>
	LTCY-320	1		10	86	115	211	0%	5%	41%	55%	100%
		2	2	25	90	69	186	1%	13%	48%	37%	100%
		3	1	22	62	66	151	1%	15%	41%	44%	100%
	<b>LTCY-320 Total</b>		<b>3</b>	<b>57</b>	<b>238</b>	<b>250</b>	<b>548</b>	<b>1%</b>	<b>10%</b>	<b>43%</b>	<b>46%</b>	<b>100%</b>
	LTCY-420	2		9	53	42	104	0%	9%	51%	40%	100%
		3			109	45	154	0%	0%	71%	29%	100%
	<b>LTCY-420 Total</b>			<b>9</b>	<b>162</b>	<b>87</b>	<b>258</b>	<b>0%</b>	<b>3%</b>	<b>63%</b>	<b>34%</b>	<b>100%</b>
	SEC-351	1	2	11	40	1	54	4%	20%	74%	2%	100%
		2	1	13	47	8	69	1%	19%	68%	12%	100%
		3	6	7	38	3	54	11%	13%	70%	6%	100%
	<b>SEC-351 Total</b>		<b>9</b>	<b>31</b>	<b>125</b>	<b>12</b>	<b>177</b>	<b>5%</b>	<b>18%</b>	<b>71%</b>	<b>7%</b>	<b>100%</b>
SEC-352	1	1		1	27	29	3%	0%	3%	93%	100%	
<b>SEC-352 Total</b>		<b>1</b>		<b>1</b>	<b>27</b>	<b>29</b>	<b>3%</b>	<b>0%</b>	<b>3%</b>	<b>93%</b>	<b>100%</b>	
<b>2 Total</b>			<b>13</b>	<b>102</b>	<b>906</b>	<b>376</b>	<b>1397</b>	<b>1%</b>	<b>7%</b>	<b>65%</b>	<b>27%</b>	<b>100%</b>
3	ELED-365	1			73	23	96	0%	0%	76%	24%	100%
	<b>ELED-365 Total</b>				<b>73</b>	<b>23</b>	<b>96</b>	<b>0%</b>	<b>0%</b>	<b>76%</b>	<b>24%</b>	<b>100%</b>
	ELED-405	2			148	37	185	0%	0%	80%	20%	100%
	<b>ELED-405 Total</b>				<b>148</b>	<b>37</b>	<b>185</b>	<b>0%</b>	<b>0%</b>	<b>80%</b>	<b>20%</b>	<b>100%</b>
	ELED-406	1			129	29	158	0%	0%	82%	18%	100%
		3			105	27	132	0%	0%	80%	20%	100%
	<b>ELED-406 Total</b>				<b>234</b>	<b>56</b>	<b>290</b>	<b>0%</b>	<b>0%</b>	<b>81%</b>	<b>19%</b>	<b>100%</b>
	ELED-407	1	3	10	97	1	111	3%	9%	87%	1%	100%
		3	1	2	55	48	106	1%	2%	52%	45%	100%
	<b>ELED-407 Total</b>		<b>4</b>	<b>12</b>	<b>152</b>	<b>49</b>	<b>217</b>	<b>2%</b>	<b>6%</b>	<b>70%</b>	<b>23%</b>	<b>100%</b>
	ELED-465	1		6	304	89	399	0%	2%	76%	22%	100%
	<b>ELED-465 Total</b>			<b>6</b>	<b>304</b>	<b>89</b>	<b>399</b>	<b>0%</b>	<b>2%</b>	<b>76%</b>	<b>22%</b>	<b>100%</b>
	MGE-481	1		33	46		79	0%	42%	58%	0%	100%
	<b>MGE-481 Total</b>			<b>33</b>	<b>46</b>		<b>79</b>	<b>0%</b>	<b>42%</b>	<b>58%</b>	<b>0%</b>	<b>100%</b>
<b>3 Total</b>			<b>4</b>	<b>51</b>	<b>957</b>	<b>254</b>	<b>1266</b>	<b>0%</b>	<b>4%</b>	<b>76%</b>	<b>20%</b>	<b>100%</b>
4	EDU-489	1		1	129	76	206	0%	0%	63%	37%	100%
	<b>EDU-489 Total</b>			<b>1</b>	<b>129</b>	<b>76</b>	<b>206</b>	<b>0%</b>	<b>0%</b>	<b>63%</b>	<b>37%</b>	<b>100%</b>
<b>4 Total</b>				<b>1</b>	<b>129</b>	<b>76</b>	<b>206</b>	<b>0%</b>	<b>0%</b>	<b>63%</b>	<b>37%</b>	<b>100%</b>
<b>Grand Total</b>			<b>37</b>	<b>186</b>	<b>3056</b>	<b>1304</b>	<b>4583</b>	<b>1%</b>	<b>4%</b>	<b>67%</b>	<b>28%</b>	<b>100%</b>

**Table 23. Teacher Work Sample Data to answer Key Question 2 (from Table 21)  
(WKU Example)**

EDU 489			Total Candidates		Total Candidates		ELED Candidates	
			Spring 2003 (n = 215)		Fall 2003 (n = 121)		Fall 2003 (n = 61)	
TWS Standards	Categories	Scores	f	Percent	f	Percent	f	Percent
<b>Holistic</b>	Exemplary	4	99	46%	47	39%	13	21%
	Proficient	3	69	32%	66	55%	47	77%
	Developing	2	29	13%	8	7%	1	2%
	Beginning	1	18	8%	0	0%	0	0%
<b>Contextual Factors:</b> The teacher uses information about the learning/teaching context and student individual differences to set learning goals, plan instruction and assess learning	Exemplary	15	92	43%	67	55%	42	69%
	Proficient	13-14	58	27%	29	24%	10	16%
	Developing	8-12	62	29%	24	20%	9	15%
	Beginning	<8	3	1%	1	1%	0	0%
<b>Learning Goals:</b> The teacher sets significant, challenging, varied and appropriate learning goals.	Exemplary	12	121	57%	88	73%	57	93%
	Proficient	10-11	75	35%	22	18%	4	7%
	Developing	6-9	17	8%	11	9%	0	0%
	Beginning	<6	1	0%	0	0%	0	0%
<b>Assessment Plan:</b> The teacher uses multiple assessments modes and approaches aligned with learning goals to assess student learning before, during and after instruction.	Exemplary	15	86	40%	69	57%	41	67%
	Proficient	12-14	62	29%	25	21%	11	18%
	Developing	8-12	66	31%	25	21%	9	15%
	Beginning	<8	1	0%	2	2%	0	0%
<b>Design for Instruction:</b> The teacher designs instruction for specific learning goals, student characteristics and needs, and learning contexts.	Exemplary	18	89	41%	75	62%	51	84%
	Proficient	15-17	89	41%	34	28%	9	15%
	Developing	9-14	37	17%	12	10%	1	2%
	Beginning	9	0	0%	0	0%	0	0%
<b>Instructional Decision-Making:</b> The teacher uses on-going analysis of student learning to make instructional decisions.	Exemplary	9	149	69%	103	85%	55	90%
	Proficient	8	28	13%	7	6%	2	3%
	Developing	5-7	36	17%	9	7%	4	7%
	Beginning	<5	2	1%	2	2%	0	0%
<b>Analysis of Student Learning:</b> The teacher uses assessment data to profile student learning and communicate information about student progress and achievement.	Exemplary	12	101	47%	86	71%	50	82%
	Proficient	10-11	67	31%	15	12%	8	13%
	Developing	6-9	42	20%	20	17%	3	5%
	Beginning	<6	5	2%	0	0%	0	0%
<b>Reflection and Self-Evaluation:</b> The teacher analyzes the relationship between his or her instruction and student learning in order to improve teaching practice.	Exemplary	15	71	33%	74	61%	46	75%
	Proficient	12-14	74	35%	25	21%	10	16%
	Developing	8-12	65	30%	22	18%	5	8%
	Beginning	<8	4	2%	0	0%	0	0%

**Table 23. Teacher Work Sample Data to answer Key Questions 3-5 (from Table 20) (WKU Example)**

**Western Kentucky University Initial Teacher Preparation Survey  
Student- (GY 02-03), First-Year (GY 01-02), & Second-Year (GY 00-01) Teachers**

<b>Level of Preparation Reported for Each Survey Question</b>						
<b>(OVERALL RESPONDENTS STTCH N = 183, FYT N = 102, SYT = 114)</b>						
<b>In general, how well do you feel WKU's teacher preparation program prepared you for actually teaching? Would you say that your teacher preparation was... ?</b>						
	<b>Excellent</b>	<b>Good</b>	<b>Fair</b>	<b>Poor</b>	<b>Don't Know</b>	<b>Mean Rating</b>
<b>St-Teacher</b>	56 ( 33% )	82 ( 48% )	30 ( 18% )	3 ( 2% )	( 0% )	3.12
<b>First Year</b>	54 ( 53% )	41 ( 41% )	6 ( 6% )	( 0% )	( 0% )	3.48
<b>Second Year</b>	27 ( 24% )	70 ( 61% )	14 ( 12% )	1 ( 1% )	2 ( 2% )	3.10
<b>WKU has adopted the following standards regarding what beginning teachers should know and be able to do. Please indicate whether you received Excellent, Good, Fair, or Poor preparation to meet each standard.</b>						
<b>1. Design units of instruction that focus specifically on the content standards for schools in Kentucky.</b>						
	<b>Excellent</b>	<b>Good</b>	<b>Fair</b>	<b>Poor</b>	<b>Don't Know</b>	<b>Mean Rating</b>
<b>St-Teacher</b>	89 ( 49% )	75 ( 41% )	18 ( 10% )	( 0% )	( 0% )	3.39
<b>First Year</b>	54 ( 53% )	41 ( 41% )	6 ( 6% )	( 0% )	( 0% )	3.48
<b>Second Year</b>	34 ( 30% )	59 ( 52% )	17 ( 15% )	3 ( 3% )	1 ( 1% )	3.10
<b>2. Use information about the community and backgrounds of your students to design learning tasks for individual students or groups of students you teach.</b>						
	<b>Excellent</b>	<b>Good</b>	<b>Fair</b>	<b>Poor</b>	<b>Don't Know</b>	<b>Mean Rating</b>
<b>St-Teacher</b>	59 ( 32% )	66 ( 36% )	48 ( 26% )	8 ( 4% )	1 ( 1% )	2.97
<b>First Year</b>	27 ( 26% )	48 ( 47% )	22 ( 22% )	5 ( 5% )	( 0% )	2.95
<b>Second Year</b>	21 ( 18% )	55 ( 48% )	36 ( 32% )	2 ( 2% )	( 0% )	2.83
<b>3. Design classroom assessments that are aligned with Kentucky content standards and tests.</b>						
	<b>Excellent</b>	<b>Good</b>	<b>Fair</b>	<b>Poor</b>	<b>Don't Know</b>	<b>Mean Rating</b>
<b>St-Teacher</b>	94 ( 52% )	74 ( 41% )	10 ( 5% )	4 ( 2% )	( 0% )	3.42
<b>First Year</b>	39 ( 39% )	50 ( 50% )	10 ( 10% )	1 ( 1% )	( 0% )	3.27
<b>Second Year</b>	34 ( 30% )	51 ( 45% )	23 ( 20% )	5 ( 4% )	1 ( 1% )	3.01
<b>4. Design and use classroom pre and post instruction assessments that reliably measure the learning results of your instruction.</b>						
	<b>Excellent</b>	<b>Good</b>	<b>Fair</b>	<b>Poor</b>	<b>Don't Know</b>	<b>Mean Rating</b>
<b>St-Teacher</b>	72 ( 39% )	75 ( 41% )	27 ( 15% )	8 ( 4% )	1 ( 1% )	3.16
<b>First Year</b>	37 ( 36% )	47 ( 46% )	15 ( 15% )	3 ( 3% )	( 0% )	3.16
<b>Second Year</b>	22 ( 19% )	57 ( 50% )	28 ( 25% )	6 ( 5% )	1 ( 1% )	2.84
<b>5. Design and use "formative" assessments to provide feedback to students and guide their learning.</b>						
	<b>Excellent</b>	<b>Good</b>	<b>Fair</b>	<b>Poor</b>	<b>Don't Know</b>	<b>Mean Rating</b>
<b>St-Teacher</b>	68 ( 37% )	76 ( 42% )	27 ( 15% )	11 ( 6% )	1 ( 1% )	3.10
<b>First Year</b>	36 ( 35% )	43 ( 42% )	22 ( 22% )	1 ( 1% )	( 0% )	3.12
<b>Second Year</b>	21 ( 18% )	61 ( 54% )	27 ( 24% )	5 ( 4% )	( 0% )	2.86

Similarly, Table 23 illustrates how student-teachers scored on the Teacher Work Sample during one semester both at the holistic level and on each of the standards that are part of the Teacher Work Sample. Thus, in spring 2003, 99 (46%) students scored 4 – Above Standard/Exemplary, 69 (32%) scored 3 – At Standard/Proficient, 29 (13%) scored 2 – Developing, and 18 (8%) scored at the Beginning Level. As the tabs on Table 23 suggest, the last columns of each page in this Excel chart allow faculty to look at how students in various programs fared as compared to the success rates of all students.

Finally, Table 24 provides the first page of a report on the response rates of a cross-section of student-, first-year, and second-year teachers to WKU's yearly New Teacher Survey. This report allows faculty to look for response patterns among teachers who took the survey during the same time period but who are at different places in their career – just graduating to two years experience teaching. Furthermore, as the tabs suggest, these data are divided across program areas.

As with all the steps in this component, it is important to remember that this is an iterative process. Again, very few institutions have used data in this way; most of us are “learning as we go.” With this in mind, in developing your timeline and anticipating your key questions, be ready to live with “good enough” in order to move to collecting your initial data. As data are collected and results are reported, it is very likely, if not inevitable, that you will discover “holes” in your data collection process and new questions to answer.

Furthermore, the same data may need to be interpreted and reported differently depending on the audience. For the sake of space, these differentiated reports are not included in this manual, but, for example, Table 22 provides the overall proficiency of all WKU candidates completing critical performances related to Kentucky New Teacher Standard 1. Although this provides important evidence of overall undergraduate program success on this standard, separate programs (i.e., elementary, middle, secondary) would need to review just the performances associated with their programs of study.

Table 23 does just that, breaking out performance on the TWS based on candidates within separate programs. But even this table could be further refined to reveal how candidates are performing on the individual indicators that make up each of the TWS standards. Although not provided here, such work has revealed that candidates may be performing well overall on a certain TWS standard and still not be as successful on a particular indicator of that standard.

Similarly, Table 24 divides respondents by program as well, thus providing rich information about how students and new teachers feel about their preparation in each program. However, this report fails to clearly connect each of the survey items to the New Teacher Standards they were designed to measure.

#### **Expected Products of this Component:**

A **completed timeline plan** that shows: (a) what program questions are assessments created to answer, (b) what performance assessment data are being collected (c) when data collection is scheduled, (d) who is responsible for data collection, analysis, and reporting, and (e) a report deadline that allows for timely decision-making. A more thorough timeline would also include additional columns, such as report format, report audience (Table 20), use of the data, and whether data are “required” (by state, institution, etc.) and “high” or “low” stakes.

**First-generation data reports** that demonstrate how results are linked to key questions regarding standards and other program/unit goals.

**Tips and Good Advice for this Component:**

1. Spend the necessary time to identify outcome related key questions and your audiences and the questions about the program that they will want answered. This will help bring focus to the assessment data collection process and help avoid wasting time and energy collecting irrelevant information.
2. Develop the timeline with the assistance of both likely data collectors and members of intended audiences to ensure that data is collected and reported in a timely manner.
3. Use the timeline as a stepping-stone towards developing an “accountability culture” in which data collection and reporting become an automatic part of the educational environment.
4. In writing reports, focus all analyses on what data say about students’ progress toward teacher standards.
5. In writing reports, be aware that the same data may be reported in various ways or different data reported depending on the intended audience.

Draft

## **Component 12:**

### **Examining the Consequential Validity of Your Assessment System**

**Key Questions:** *What is the consequential validity of your assessment system in terms of its effect on the qualifications of your graduates and their success as educators? What are the consequences of your efforts to use your assessment data to make program improvements? Are there any unintended adverse consequences?*

#### **Task for this Component:**

Evaluate the positive and adverse consequences of your assessment system on your graduates. Conduct studies to determine whether performance levels of program graduates predict subsequent job success using multiple criteria. Conduct studies to determine candidate characteristics that are related to success on your assessments. Conduct studies to determine how performances on your assessments are related to each other. Evaluate the pattern of your results to determine whether any groups of candidates score better or worse due to biasing factors that are not related to their success as educators. Also, gather data and conduct studies to determine whether changes made on the basis of your assessment system produce the intended program improvements and whether there are any adverse consequences. Because the research methods and designs for these studies will vary, the steps described in this chapter are merely illustrative of the kinds of steps that should be taken. Evaluation of the validity of the interpretations resulting from your assessment system will necessarily require an ongoing conversation among the members of your professional community and with the larger constituencies it serves.

#### **Rationale for this Component:**

Beyond the technical quality of your assessments, your ability to summarize candidate performance levels, and your ability to generate useful reports, your assessment system should maximize positive professional and societal effects and minimize adverse ones. The consequential aspect of validity is concerned with the value implications and social consequences of interpretations made on the basis of your assessment system (Messick, 1980, 1989, 1995). These concerns include the following:

- Is the professional community being better served? Are the families and schools of your state being better served by your graduates? Your assessment system should also convey the right messages about what it means to be a highly qualified educator. Decisions made on the basis of the assessment results should not inappropriately exclude persons from becoming educators or inappropriately eliminate them from opportunities for advancement. It is also important to study your program improvement efforts to insure your changes produce the intended benefits and there are no unintended adverse consequences (e.g., undue restriction of the curriculum, undue restrictions on academic freedom and other rights of your faculty and students).
- Does our assessment system value diversity in teachers and in teaching? Hence, it is important to gather evidence showing that the results of using your assessment system are positive and that the adverse consequences are minimal and they are not due to factors unrelated to your candidates' success as educators.
- Are there any unfair disadvantages for some candidates? Does your assessment system reveal the actual abilities of your candidates? High performance should mean that standards have been met, but the causes of low performance could be multiple and due to factors other than whether or not the candidate is capable of meeting the standards. Hence,

it is always important to take into account other relevant information when making “high-stakes” decisions. All decisions made on the basis of assessment data have some degree of uncertainty attached to them. Therefore, have allowances been made for your candidates to appeal the decisions made on the basis of your assessment system?

A harder question to address is: Does the assessment system as a whole have an impoverished view of competence such that the potential teaching (or leadership) qualities of some groups of persons are discounted or marginalized? Compromises were made in the development of your assessment system reflecting the values of the participants and their willingness to accept definitions of evidence as “good enough” and cut-off scores as consequential. Do we really know what “good” teaching (or “good” leadership) looks like in all of its manifestations—even for the narrow goals set for public schooling in the United States of America? Your assessment system only represents the current consensus of your professional community about factors that you presently believe are warranted as important to success as an educator. However, the consequences of our decisions made on the basis of our assessment systems must be continually questioned in light of new evidence and changes to our profession.

**Process for this Component:**

Because the process steps will vary according to the purpose and design of the study, the steps described below are appropriate for only one type of study that should be undertaken as you evaluated the consequences of decisions made on the basis of your performance assessments. The steps below are for a study of the fairness of a performance assessment. The steps are illustrated using the Renaissance Teacher Work Sample.

**Step 1.** Identify important demographic characteristics of your candidates, such as gender, race/ethnicity, age, linguistic background, and teaching major. This information should be available from your institutional student information database, if it has not also been entered into your colleges’ database on your teacher candidates.

**Step 2.** Determine a time frame for your study, such as all teacher education graduates from spring 2002 through spring 2004.

**Step 3.** Crosstabulate the categories of each demographic variable with the categories created by the consequential cut-scores on your performance assessment. For example, score ranges on the Renaissance TWS can be converted into the developmental categories of 1 = Beginning, 2 = Developing, 3 = Proficient, and 4 = Expert. Inspections can then be made by demographic category of the proportion of your teacher candidates who fall into each of the consequential categories of your performance assessment. If your sample size is sufficiently large to include multiple representatives in each category, then a Chi-Square test of association could be used to determine whether there is a statistically significant association between category membership on the demographic variable and the consequential decision levels of your performance assessment. Ideally, you would not find a statistically significant association. With sufficient sample size, mean score differences by demographic category on your performance assessment could also be examined using analysis of variance. Table 25 shows an example for TWS performances levels at Western Kentucky University by the sex of the teacher education candidates.

**Step 4.** Investigate thoroughly the causes of any consistent pattern of association or mean difference. Additional statistical analyses might be conducted to determine whether the pattern is due to other predictors of candidate success. For example, a mean difference in TWS performance levels between male and female teacher candidates might not indicate gender

bias, if the lower on average TWS performances of the males were related to their having lower average Praxis-I scores, lower average ACT scores, and lower admission grade point averages than the females.

**Expected Product of this Component:** Reports of consequential validity studies regarding the presence or absence of unintended bias in the assessment or scoring process.

**Tips and Good Advice for this Component:**

1. Plan ahead and determine in advance all important candidate demographic characteristics to be include in your college's database.
2. It may take several years to collect data on a sufficient number of teacher candidates to be able to conduct statistical analyses and draw meaningful conclusions.

**Table 25. Crosstabulation of WKU TWS Scores by Sex\***

	<i>TWS Overall Score Categories (N and Row %)</i>			
	<b>Beginning</b>	<b>Developing</b>	<b>Proficient</b>	<b>Exemplary</b>
<b>Male (N=51)</b>	1 (2%)	9 (18%)	30 (59%)	11 (22%)
<b>Female (N=161)</b>	3 (2%)	16 (10%)	85 (53%)	57 (35%)

\* $\chi^2 = 4.525$  (n.s.)

## References

- American Educational Research Association (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Crocker, L. (1997). Assessing content representativeness of performance assessment exercises. *Applied Measurement in Education, 10*, 83-95.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Fort Worth, TX: Harcourt, Brace, Jovanovich.
- Danielson, C. (1996). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Danielson, C. (1997). *Performance tasks and rubrics: Upper elementary school mathematics*. Larchmont, NY: Eye on Education.
- Danielson, C., McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Denner, P. R., Salzman, S. A., & Bangert, A. W. (2001). Linking teacher assessment to student performance: A benchmarking, generalizability, and validity study of the use of teacher work samples. *Journal of Personnel Evaluation in Teacher Education, 15*(4), 287-307.
- Denner, P. R., Salzman, S. A., Newsome, J. D., & Birdsong, J. R. (2003). Teacher work sample assessment: Validity and generalizability of performances across occasions of development. *Journal for Effective Schools, 2*(1), 29-48.
- Dunbar, S. B., Koretz, D., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*, 289-302.
- Lunz, M. E., & Schumacker, R. E. (1997). Scoring and analysis of performance examinations: A comparison of methods and interpretations. *Journal of Outcome Measurement, 1*(3), 219-238.

- McConney, A., & Ayres, R. R. (1998). Assessing student teachers' assessments. *Journal of Teacher Education*, 49(2), 140-150.
- National Council for Accreditation of Teacher Education. (2000). *NCATE 2000 Unit Standards*. Washington, DC: Author.
- Renaissance Partnership for Improving Teacher Quality. (2001). *Renaissance Teacher work sample: Teaching process standards, and scoring rubrics*. Western Kentucky University, Bowling Green, Kentucky. Web site: <http://fp.uni.edu/itq/default.htm>
- Schalock, M. (1998). Accountability, student learning, and the preparation and licensure of teachers: Oregon's teacher work sample methodology. *Journal of Personnel Evaluation in Education*, 12, 269-285.
- Schalock, H. D., Schalock, M., & Girod, G. (1997). Teacher work sample methodology as used at Western Oregon State College. In J. Millman (Ed.). *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 15 - 45). Thousand Oakes, CA: Corwin Press.
- Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, 30(3), 215-232.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.